TABLE OF CONTENTS

I. INTRODUCTION

The packet radio project relies heavily on station software for a variety of control, coordination and monitoring functions. The role of BBN in developing this software is to specify, design, implement and deliver programs which implement these functions.

Progress in the previous quarter brought packet radio station software to the brink of functional operation; this quarter has seen the promise of that and prior support efforts fulfilled. The station is now a functioning entity, complete with all software and support tools required for delivery of the first version to SRI, the contractor coordinating the packet radio project. As this quarter closes, the initial efforts to run the complete station software on SRI equipment are being made. The familiarization and transfer activities of the next quarter will be a major cooperative accomplishment in the packet radio project as a whole. The work during this past quarter has prepared for and ensured the success of that step.

Most of the sections below describe particular modules, or processes, which are specific parts of the station software. Each has its own individual purpose and responsibilities. The separability of these processes has reduced interdependence and enabled a parallel implementation effort, with the result that all modules have reached operational status during this quarter.

In addition to software preparation, BBN continues to contribute to progress of the packet radio project in areas of design study and support development. This is detailed in following sections dealing with publications and support software. We also continue an active role to design, deploy and debug hardware, although only minimal needs for such effort have arisen during this quarter, as described below in the section on hardware.

II. MEETINGS, PUBLICATIONS, AND DEMONSTRATIONS

This quarter opened with a meeting at Collins Radio to discuss SPP protocol. BBN attendees reached agreement with representatives of other contractors on outstanding issues. As the successful conclusion of this meeting, the SPP protocol was frozen. BBN issued a statement of the resolution of each particular item.

On April 6 BBN hosted a meeting with SRI personnel. Progress was reviewed and methods for future cooperative efforts were discussed. This discussion was instrumental in impressing upon SRI the need for familiarization activities begun later this quarter.

On May 1 BBN hosted a meeting of researchers and developers of the satellite network. The packet radio project obtains synergistic benefit from overlapping effort between the two projects, since both require gateways and internetwork (TCP) facilities. At this meeting several packet radio network capabilities were demonstrated, as described below.

The final meeting of this quarter was held at SRI on May 3-4 and addressed measurement issues. As a result of this meeting, BBN agreed to prepare a working paper on the preliminary specification of the station's measurement process.

BBN published two PRTNs this quarter. PRTN 174, "Packet Radio Network Station Labeling Process," describes and delimits the functions performed by the first version of the control process. PRTN 177, "SPP Definition," provides a precise definition of SPP

protocol, complete with a state diagram and state transition table.

BBN prepared a summary of problems with PR software. This annotated list was sent to Collins and SRI, and a detailed discussion of one unresolved problem was sent to Collins. Our continued effort to identify and, where possible, propose corrections for dysfunctional PR operation is an important part of our role in coordinating our station development work with work by other contractors on other packet radio components.

SRI published PRTN 167, which presents several ingenious ways to extract measurements from search packets the station receives. It was flawed, however, by erroneous mathematics. We circulated a development of the correct formulae.

During this quarter various demonstrations of station software were performed. In April, SPP protocol was tested between a test program in the station and the station PR. Connections were opened from either end; information (text) packets were exchanged and printed; connections were closed; and display memory ("DM") executable control packets were sent to the PR and their responses received in the station. Later in April the forwarding of intranet packets from a level 1 PR to a level 2 PR through the station was accomplished. These achievements set the stage for a major demonstration prepared for visitors from other organizations.

At the May 1 meeting reported above, a demonstration was held

which illustrated the functioning of several modules central to the Packet Radio station or terminal.

The first configuration demonstrated uses the user TELNET and TCP0 developed by Dr. Cerf's group at Stanford University in their LSI-11. The LSI-11 represents a typical, single PR net TCP terminal. It is connected to its PR, which is in turn connected to the station PR by a cable. The cable between PRs represents an arbitrary radio routing of the terminal's traffic through the PR net. Traffic entering the station from its PR flows through the connection process to the gateway process and out the IMP11A interface to the ARPANET. It is addressed to a service host on the ARPANET, namely a TENEX which is running a TCP and a server TELNET. Traffic from the server TELNET takes the reverse path back to the terminal.

The second configuration demonstrated uses the local terminal on a PR to obtain a connection, as in the previous configuration, to an ARPANET service host. In this case, however, two PDP-11s are used. Traffic originating at the PR local keyboard goes through the PR (number 2), into the attached PDP-11 which is running a connection process. The connection process forwards it by readdressing it to what is essentially a service host on the PR net (supplying TCP and gateway service). The traffic is sent back to PR number 2 and through it to PR number 1, through PR number 1 to the PR net service host PDP-11, and through a connection process in that PDP-11 to an SPP user TELNET. The TELNET hands the traffic to a TCP, which in turn hands the (now TCP protocol with internetwork headers) traffic to a gateway, still in the PR net service host. The traffic leaves the gateway, traverses the ARPANET and TENEX service host's TCP and arrives at the server TELNET as in the first configuration. Again, traffic from the server TELNET back to the terminal takes the reverse path.

This demonstrates a considerable degree of accomplishment, since the connection process (using a wild connection in the first configuration), gateway, TCP in a PDP-11 (useful for complicated terminals, terminal multiplexors, and TCP service hosts if needed), a user TELNET which can communicate at the SPP level with PR net terminals, TCP and server TELNET services on the ARPANET, and, last but not least, SU's LSI-11 implementation of TCP and user TELNET in a prototype terminal are all functioning.

In conclusion, an additional demonstration -- this time informal -- was performed by BBN personnel who remained at SRI after the May 3-4 meeting there. Station software labeled several PRs for the first time over the radio channel. This also marked the beginning of procedures to familiarize SRI personnel with station operation.

III. CONNECTION AND GATEWAY PROCESSES

At the beginning of the quarter, the connection process and gateway process had been written and debugging had begun. At that time we were using the link test support program in the Packet Radio Unit (PRU) to loop packets back to the station.

Work on debugging the connection and gateway processes continued in this quarter. The connection process implements the station-PR protocol (SPP) to support communications between applications processes in the station and the PRUs. This quarter, with the receipt of Collins CAP software, which implements SPP in the PRUs, we were able to debug the connection process's interaction with the PRU software. At this time, several differences in the interpretation of the SPP protocol in the two implementations became apparent and were resolved.

Use of the connection process both for forwarding packets between PRUs, and for supporting communications between the gateway process in the station and the PRUs was demonstrated at BBN on May 1. By the end of the quarter, we had also designed and coded the interface between the connection and station control processes and had used these two processes in tests both at BBN and SRI to label small PR nets.

With the receipt of Collins CAP software and the completion of the connection process, we were able to debug the gateway process acting as a gateway between the ARPANET and the PR net. The gateway

consists of interfaces to the connection process and XNCP, and a set of gateway routines which interpret internet headers to re-address packets received from either network to the destination network. In addition, the gateway interfaces to a TCP implemented in the station. In the May 1 demo, operation of the gateway was demonstrated in two ways. First, packets from a terminal on the PR net were directed to a TCP Telnet in the station. The station TCP imbedded the packets from the terminal in internet packets, then sent these packets to the gateway which re-addressed them to their destination, a TCP implemented in a BBN TENEX system on the ARPANET. Similarly, packets received from the TCP on the ARPANET were routed by the gateway to the TCP in the station. In a second test, internet packets were generated on the PR net by a terminal using TCP0 implemented in an LSI-11. These packets were addressed to the station gateway which re-addressed them to the TCP on the ARPANET. Packets from the ARPANET TCP were received by the gateway and re-addressed to the TCP0 on the PR net.

By the end of the quarter, both the connection and gateway processes have undergone considerable testing in the BBN PR net and we are looking forward to continued testing in the SRI PR net prior to delivering this software to SRI.

IV. CONTROL PROCESS

The control process in the station is responsible for labeling (determining how packets are to be routed through) the network. This quarter we designed, implemented, and debugged the initial version of the control process.

A. Functional Design

The initial functional design was described in PRTN 174, "Packet Radio Network Station Labeling Process", issued in March. In this design, the process uses ROPs as its only source of information on the existence, labeling, and connectivity of PRs. Based on these, it performs network initialization, labeling PRs which are unlabeled or have the wrong label (as might happen if the station is reinitialized). This first design doesn't reroute around failed repeaters, track mobile terminals, or readjust routing to improve traffic flow. The design takes into account some properties of the packet radio network -- that an acknowledgement may be lost although the packet it acknowledges was received, and that packets may not be received in the same order in which they were sent -- by using timeouts (based on the concept of a maximum packet lifetime).

B. Connection Process Interface

The station connection process must be able to find out what routes are currently assigned so it can route packets received for forwarding or originated by other station processes. Two different

interfaces between the control and connection processes were implemented. In the first, the two processes resided in the same ELF address space, and the connection process accessed the control process route table directly, using locks to avoid conflict. However, as implementation progressed, it turned out that the two processes would not fit into one address space. Thus the second interface was implemented, in which the connection process keeps its own copy of routing information, and the control process communicates route changes to it via an interprocess port.

C. Operator Interaction

The program interacts with the operator in several ways:

During initialization, the operator may define the format to be used for routes in packets -- i.e. how may bits are used for each level.

At any time, the operator may interrupt the program to display the currently assigned labeling and the current network connectivity.

The operator may also define non-PR devices (terminals) and give the correspondence between them and their PRs. There is as yet no way for the station to establish this correspondence, which is essential for forwarding terminal traffic, automatically.

For network testing purposes, the program may be placed in a manual mode, in which it does not automatically assign labels.

Instead, the operator enters the desired labeling and the program sends out only those values.

D. Implementation

The initial version of the program was implemented, and debugged at BBN. Testing in the BBN setup, however, with only two PRs and no radio links, is necessarily limited. A brief test was performed at SRI, with five PRs communicating by radio. More testing will be done at SRI in the coming quarter, preparatory to delivering the software to SRI.

V. CROSS-RADIO DEBUGGER

The cross radio debugger, XRAY, has been completed. Testing and use have discovered a few bugs, which were fixed, and permitted several small design improvements which make the debugger easier to use.

As an experimental feature, another command has been implemented which was not anticipated in the initial design. This is the ability to load overlay modules into PRs. Any of the overlays supplied by Collins (INFOP, for text communication; PARAM, for parameter setting; and CNTRL, for cross-PR debugging) may be loaded. Each module was extracted from tape supplied by Collins, reformatted, and segmented as arguments to alter memory ("AM") packets.

XRAY has proven useful in adjustment and diagnosis of PR operation in connection with testing other portions of station software. Publication of a manual and full commenting of the source code will complete the XRAY development task. We plan to evaluate possible enhancements to XRAY as the need for them arises. A trimmed down version of XRAY (with the overlay load command excised) has been prepared which fits in the restricted memory available on the SRI machine.

VI. TCP DEVELOPMENT

A. PDP-11 TCP

Using the timing facilities installed in ELF and the PDP-11 TCP which were reported last quarter, the performance of the TCP was investigated in some detail. An interesting outcome of this investigation was the observation that 50 percent of the total time consumed by all processes in the station while the TCP was in operation was spent executing ELF primitives for the TCP. 25 percent was spent in processes other than TCP processes (e.g. XNCP and IOX) and that the remaining 25 percent was spent in the TCP proper.

In an effort to improve the TCP performance, the user interface to the TCP was modified to reduce the number of ELF primitives executed (principally VMOVs). This resulted in a significant but not impressive improvement.

A rudimentary user TELNET was written to operate with the TCP. In its current form, it allows terminals on the PR net to connect via SPP to the user TELNET and from there, to connect to host computers via TCP. This will be augmented to permit terminals on the PDP-11 itself to also connect to host computers via TCP. In this form, it may be the heart of a terminal concentrator or multiplexor.

VII. SUPPORT SOFTWARE

A. ELF and the BCPL Library

Substantial improvement and expansion of the ELF BCPL library was done during this quarter. Most of this effort was aimed at reducing the memory space requirements of the packet radio station by placing many of the utility routines which were previously not part of the library into the library and arranging for the sharing of the library amongst the various station processes.

The ELF BCPL library is shared by placing it in a specific (7) physical memory page. When ELF is started, the initialization procedure maps that physical page into a virtual storage map and examines its contents to verify the integrity of the shared library. If all is well, a user EMT is defined to execute in that address space. This EMT is executed by each process which wishes to share the library. The net effect of executing this EMT is to map the library in page 7 of that process' address space.

A special program is responsible for installing a new library. This program is included as part of the shared library and contains the code for the user EMT mentioned above. It also contains code to copy the new library into physical page 7 and then restart ELF.

Entry to the various library routines is accomplished through a transfer vector at the beginning of the library page. A dummy library file is loaded with the user's program to define the location of each routine's entry. Since BCPL normally uses a level

of indirection to call a subroutine, there is no extra cost associated with the use of the shared library. The use of the entry vector permits the library to be changed without requiring the user's program to be relinked.

To maximize the space-saving of the shared library, as many utility routines as will fit into one page must be included. Therefore, the dynamic storage allocator and delayed signalling routines from the TCP were adapted to be library routines and included. The adaptations were necessary to generalize the routines such that they could be used for programs other than the TCP.

The dynamic storage allocator permits allocation of arbitrarily sized blocks of storage. The available storage is divided into various pools and the allocation can be made out of a particular pool. The storage in a pool is, in a sense, reserved for that pool and not available to another pool. A shared pool can be used if the storage reserved for a particular pool is exhausted. This arrangement prevents the entire dynamic storage area from being allocated for a particular class of use to the exclusion of another class of use.

The delayed signalling routines extend the ELF interprocess signalling and timing primitives so that an arbitrary process can be sent a particular signal (an ELF signal consists of 32 bits of information of which 8 identify the signaller) after an arbitrary (2^31-1 millisecond) delay (ELF timing primitives restrict the delay to 2^16-1 milliseconds).

The process of debugging the shared library turned up a number of bugs in the ELF kernel, including a piece of code which had never been executed at all. Most of these bugs were related to the manipulation of storage maps. As a consequence, the installation of the shared library also resulted in the improvement of ELF itself.

B. XNET -- Cross-net Debugger

The cross-net debugger has undergone several improvements. Some of these were in the nature of glitch and bug fixes. Others were additional commands to streamline often repeated procedures such as the establishment and loading of ELF processes. The XNET manual was rewritten to include all the added features and to correct errors.

The reporting of traps was considerably improved. Previously, the only indication given when a process trapped or halted was that it had halted. Now, the error code for the trap and the interpretation of that error code is given as well as the PID of the process which initiated the trap. In conjunction with the correct reporting of a process' status after a trap or halt, this information gives the programmer an immediate and accurate indication of how his program has malfunctioned.

The cross-net IO (XIO) feature reported earlier has been expanded to include cross-net input as well as output. This means that programs which do terminal IO can now be completely debugged cross-net. Previously, programs which needed terminal input had to

be debugged from a location which had a terminal connected directly to the PDP-11.

Hexadecimal input and output has been added to XNET making it possible to enter and print numbers in forms compatible with those used by other components of the PR network such as the PROM operating system in the PR.

C. 11BCPL Compiler

Very little work was necessary on the 11BCPL compiler. A few bugs were reported and corrected. The code generator was improved to combine successive bit tests into a single bit test where possible. The most recent compiler was installed on all BBN TENEX systems.

VIII. HARDWARE

During this quarter the BBN packet radio equipment was moved to our new building. This provides adequate space and a better environment for computer-related equipment. In addition to installation of the equipment in the limited access "North Bay," we have provided lines to a room on the floor where our offices are located. The consoles of all packet radio equipment, as well as some TENEX terminals, are located in this convenient setting. Only rarely is it necessary to leave this area and enter the North Bay -- to restart a crashed PDP-11 or PR, or to run diagnostics.

Both digital units (PRDUs) were converted to 300/1200 baud EIA operation. This principally involved returning PROM boards to Collins for reprogramming. The change supported the conversion to a TI 733 terminal with tape cassettes as the PR local terminals. This allows faster loading of software supplied by Collins. We also prepared a tape of patches to Collins software which loads via the 733.

Besides the PROM boards, two RAM boards were returned to Collins for installation of additional memory needed for the latest PR software.

Although its delivery was somewhat later than expected, the PDP-11 for use as a second station and for terminal multiplexor development arrived during this quarter. It is undergoing checkout as the quarter closes; no major problems are anticipated.

During this quarter discussions have arisen regarding the need for increased station hardware to support measurement requirements. We anticipate quantifying this need and taking management action in the next quarter to procure the necessary equipment.

TABLE OF CONTENTS

I. INTRODUCTION

In the last quarter, we developed spectral warping techniques and applied them to linear predictive vocoding, with the objective of either improving the vocoded speech quality for a given bit rate or lowering the bit rate for a given speech quality.

In the area of objective speech quality evaluation, we developed a number of perceptually consistent spectral distance measures for the purpose of determining an objective error over a given frame of synthesized speech.

Our work on real-time LPC implementation was directed towards bringing up an operating system for our SPS-41/PDP-11 facility.

In speech quality evaluation, we have run some pilot studies impinging on the design of our next study of subjective quality. A provisional design of the study is now complete, and we are in the process of generating the stimuli for it.

II. PUBLICATIONS AND MEETINGS

During the last quarter, we issued 4 ARPA NSC Notes 86-89. NSC Note 86 deals with a general analysis-synthesis scheme for performing arbitrary spectral warping (or distortion) of speech signals without the need for pitch extraction, while NSC Note 88 considers the application of spectral warping in linear predictive vocoding. A general framework that we have formulated for objective evaluation of vocoder speech quality is presented in NSC Note 87. In NSC Note 89, we introduce the notion of perceptual consistency for spectral distance measures, and give examples of perceptually consistent measures. All these 4 NSC Notes were also presented, along with the paper on lattice covariance methods (NSC Note 75), at the 1976 IEEE International Conference on Acoustics, Speech, and Signal Processing (Philadelphia, April 12-14).

At the March ARPA meeting held at ISI, we presented the specifications for ARPA LPC System II (NSC Note 82), and discussed the effects of lost packets on speech intelligibility (NSC Note 78).

III. APPLICATION OF SPECTRAL WARPING IN LINEAR PREDICTION

In ordinary linear prediction the speech spectral envelope is modeled by an all-pole spectrum. The error criterion employed guarantees a uniform fit across the whole frequency range. However, we know from speech perception studies that low frequencies are more important than high frequencies for perception. Therefore, a minimally redundant model would strive to achieve a uniform perceptual fit across the spectrum, which means that it should be able to represent low frequencies more accurately than high frequencies. In an attempt to achieve such a uniform perceptual fit, we applied our recently developed linear predictive spectral warping technique to LPC vocoding. The details of the resulting vocoder that we denote by the initials LPCW are contained in NSC Note 87, which is included in this report as Appendix A.

Synthesis experiments performed using the LPCW vocoder indicated that the introduction of spectral warping produced a saving of about 10-15% in bit rate without affecting the speech quality.

IV. OBJECTIVE EVALUATION OF SPEECH QUALITY

A. A General Framework

We proposed a framework within which we have begun a step-by-step program to develop objective measures for vocoded speech quality that are consistent with results from subjective tests. Details of this framework are given in NSC Note 86, which is reproduced in this report as Appendix B. The first part of this step-by-step program, described below, deals with the development of a suitable objective measure for computing the error between spectra of synthesized speech and natural speech for a given frame.

B. Perceptual Consistency of Spectral Distance Measures

Given two smoothed short-time speech spectra, a fundamental problem in speech processing is to determine the distance or the amount of deviation between the two spectra. Considering speech compression in particular, this problem is encountered in a variety of situations such as (1) spectral sensitivity analysis (needed for optimal parameter quantization), (2) variable frame rate transmission, and (3) objective speech quality evaluation. In the context of (3), we have used some simple distance measures but the results we obtained did not correlate well with subjective scores (see Appendix B for details). Motivated to develop better distance measures in this application, we first introduced a notion of perceptual consistency for spectral distance measures, based on previously known subjective results

on formant frequency difference limen. We then developed a class of perceptually consistent distance measures by suitably defining the error at each frequency between the given two spectra and by proper frequency-weighted averaging of this error. The definition of perceptual consistency and the description of a class of perceptually consistent measures according to this definition are given in NSC Note 89, which is reproduced in this report as Appendix C. Below we describe the work performed on objective measures since the time NSC Note 89 was written.

1. A Smooth Frequency Weighting Function

The frequency weighting function $A(\omega)$ obtained from the work of Stevens has "corners" or places of sharp changes, where the first derivative of $A(\omega)$ is discontinuous (see Fig. 2 in Appendix C). With this weighting function, the plots of spectral distance versus formant frequency shift do not sometimes exhibit the required asymmetrical patterns when the formant frequency being shifted is quite close to one of these corners (especially the first one at 400 Hz). We resolved this problem by employing a smooth approximation below to the original $A(\omega)$, which is given below:

$$A'(\omega) = \frac{\omega^3[\{1-(\omega/\omega_1)^2\}^2 + (2c\omega/\omega_1)^2]}{(1 + \omega/\omega_2)^{11}} \qquad (1)$$

where ω_1=1300 Hz, ω_2=2500 Hz, and c=.6. The plots of $A(\omega)$ and its smooth approximation $A'(\omega)$ are shown in Fig. 1. Henceforth, we shall use $A'(\omega)$ for frequency weighting of spectral error and

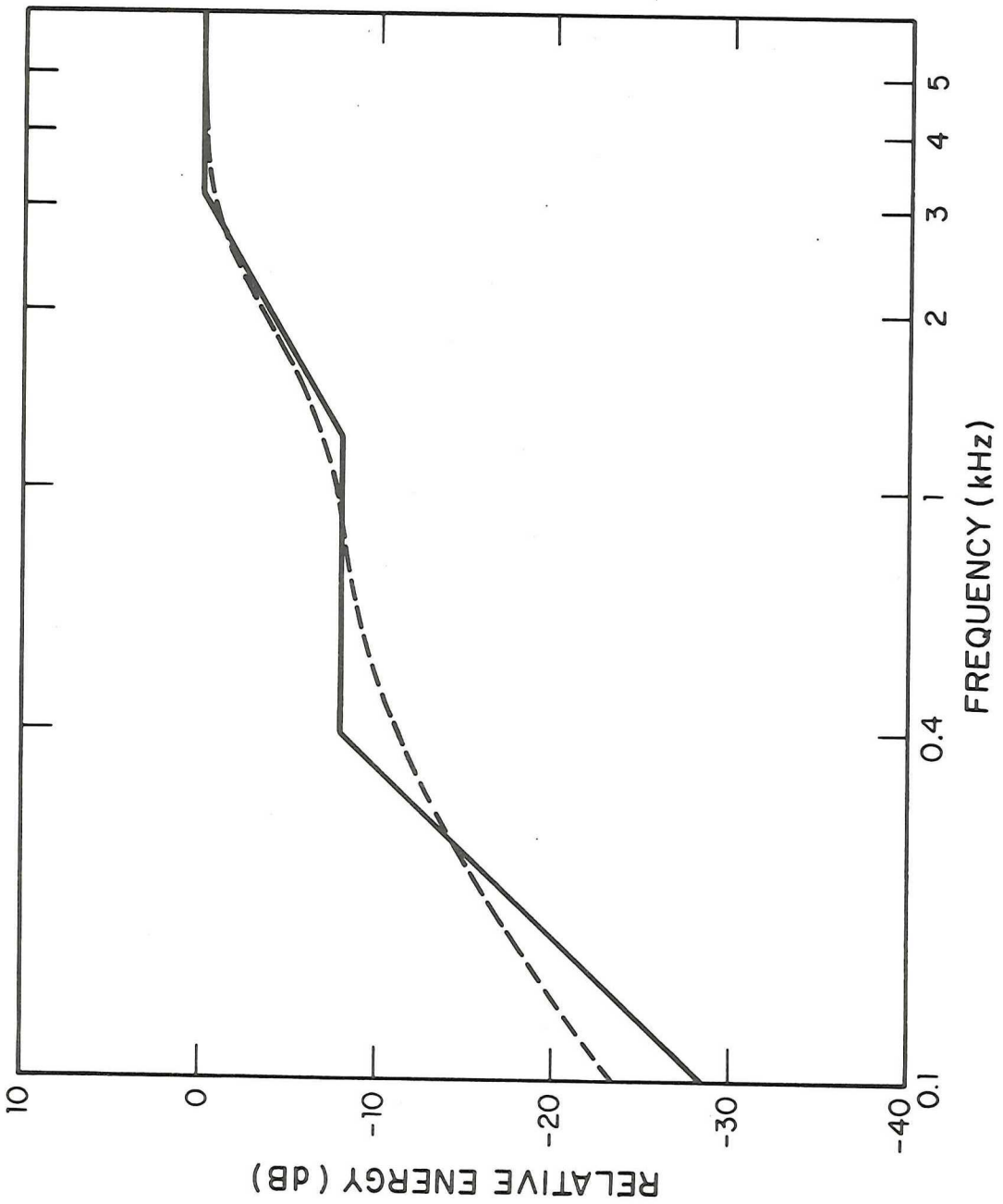


Fig. 1 Plots of weighting functions A(ω) (solid line) and A'(ω) (dashed line)

refer to it simply as $A(\omega)$.

2. Digital versus Analog Spectra

Given a set of N formant frequencies and bandwidths $\{F_i, B_i, i=1, 2,\ldots,N\}$, the power spectrum can be computed either in the digital domain at frequencies along the unit circle in the z-plane, or in the analog domain at frequencies along the $j\omega$-axis in the s-plane, as follows:

Digital Spectrum:

$$P_d(\omega) = \frac{G^2}{\prod_{k=1}^{N} \left| 1 + a_{1k}z^{-1} + a_{2k}z^{-2} \right|^2_{z=e^{j\omega}}} , \tag{2a}$$

where

$$a_{1k} = -2\,\rho_k \cos(2\pi F_k T) , \tag{2b}$$

$$a_{2k} = \rho_k^2 , \tag{2c}$$

$$\rho_k = \exp(-\pi B_k T) , \tag{2d}$$

and T is the sampling period in seconds.

Analog Spectrum:

$$P_a(\omega) = \frac{G^2}{\prod_{k=1}^{N} \left| s^2 - 2\sigma_k s + \sigma_k^2 + \omega_k^2 \right|^2_{s=j\omega}} , \tag{3a}$$

where

$$\sigma_k = \pi B_k , \tag{3b}$$

$$\omega_k = 2\pi F_k . \tag{3c}$$

The gain term G^2 in (2a) or (3a) is determined by the type of normalization used (see Appendix C). A major difference between the two spectra is that, unlike the analog spectrum, the digital spectrum is periodic with a period equal to half the sampling frequency, 1/2T. Also, a peculiarity of the digital spectrum introduced by (2b) (and the periodicity property) can best be illustrated by considering a single-formant spectrum (N=1). The two digital spectra with single formants at $1/4T+\Delta F$ and $1/4T-\Delta F$, with the same bandwidth, are mirror images of each other about a frequency of 1/4T. In the special case of a formant at 1/4T exactly, the digital spectrum is therefore even-symmetric about 1/4T. Suffice it to say, therefore, that digital and analog spectra have different properties.

The results presented in Appendix C (specifically, the necessary condition for perceptual consistency, and the plots in Figs. 1,3-6) were obtained employing digital spectra in the computation of spectral distance. But, synthetic speech stimuli used in Flanagan's subjective tests (Ref. 10 in Appendix C) were prepared using an analog formant synthesizer. Since we have used Flanagan's results as a basis for defining perceptual consistency, we repeated the work reported in Appendix C but using analog spectra instead of digital spectra. We found that the necessary condition for perceptual consistency stated in Appendix C continues to hold, and that the new plots of spectral distance versus formant frequency shift were very similar to the ones depicted in Figs. 1,3-6 of Appendix C.

The form of Flanagan's cascade synthesizer suggested a new normalization technique that we shall call DC normalization. This normalization procedure is explained below.

3. DC Normalization

In Appendix C, we mainly considered two methods of normalizing the given spectra $P_1(\omega)$ and $P_2(\omega)$ before computing the distance or deviation between them. Those were: 1) Arithmetic mean (AM) normalization, where the averages of the two spectra are made equal, and 2) Geometric mean (GM) normalization, where the averages of the log spectra are made equal. A third normalization method that we have investigated results in the two spectra having the same value at DC (i.e., at $\omega=0$ or $z=1$). We have considered this common DC value to be unity, motivated by the physical reasoning that the volume velocity component at DC encounters a unity gain when passing through the vocal tract.

Employing DC normalization with digital spectra, the gain term G^2 in (2a) becomes

$$G^2 = \prod_{k=1}^{N} (1+a_{1k}+a_{2k})^2. \tag{4}$$

Similarly, with analog spectra, we obtain (see (3a)):

$$G^2 = \prod_{k=1}^{N} (\sigma_k^2 + \omega_k^2)^2 \tag{5}$$

It can be shown that the necessary condition for perceptual consistency stated in Appendix C for GM-normalized spectra also holds true for DC-normalized spectra. This means that any spectral distance measure between two DC-normalized spectra, which is a function only of the ratio of the two spectra, cannot be perceptually consistent. Therefore, to achieve perceptual consistency with DC normalization, we exclusively considered the error definition in the linear spectral domain given by (2) in Appendix C, and investigated the use of a number of weighting functions described there. The important result of our investigation is that the class of distance measures, given by (6) in Appendix C, between two DC-normalized spectra $P_1(\omega)$ and $P_2(\omega)$, are perceptually consistent if we employ a linear spectral error and a loudness weighting of $[P_1(\omega)A(\omega)]^{1/3}$ or $[P_2(\omega)A(\omega)]^{1/3}$. (In Appendix C we showed that for GM normalization, perceptual consistency is achieved with a linear spectral error and with weighting by $A(\omega)$.)

C. "Scaling" of Spectral Distance Measures

For perceptual consistency, we required the spectral distance as a function of formant frequency shift to exhibit roughly the same shape as was obtained in Flanagan's subjective experiment. We did not place any condition on the actual magnitude of the spectral distance. There are situations, however, where we have to compare the magnitudes of spectral distance obtained for different analysis conditions. Specifically considering the objective quality evaluation

problem, spectral distances computed for different speech sounds need to be compared. Such comparisons can be meaningfully done only after suitably scaling the distance values.

Proposed Criterion for Scaling: A reasonable criterion for scaling is to require the spectral distances due to formant frequency shifts equal to the difference limen (DL) at each frequency to be approximately the same for different formant frequency values.

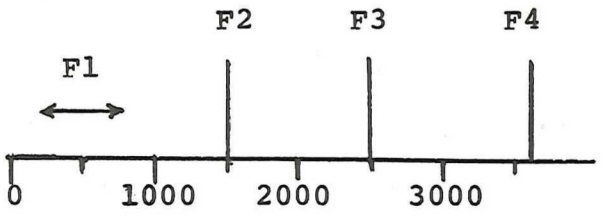
The scaling criterion stated above may be regarded as part of the requirements for perceptual consistency of a spectral distance measure.

To evaluate any given scaling method, we took Flanagan's DL values for the first two formant frequencies F1 and F2 given in Table 1, and computed spectral distances for those 12 conditions of formant frequency shift. As figures of merit, we then computed the following two quantities from those 12 distance values: 1) Maximum to minimum distance ratio, R1; 2) Arithmetic mean to geometric mean ratio of the 12 distance values, R2. Both R1 and R2 should be close to unity to satisfy the proposed scaling criterion.

For experiments involving scaling, we mainly considered the perceptually consistent measures that we developed. They employ linear spectral error definition, and either GM normalization and $A(\omega)$ weighting or DC normalization and loudness weighting $L(\omega)$. For frequency averaging, we used the absolute spectral error

DL for F1:

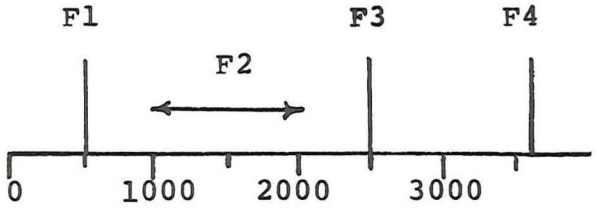
F1	-DL	+DL
300	17	12
500	25	27
700	27	19



(a)

DL for F2:

F2	-DL	+DL
1000	20	50
1500	45	75
2000	90	20



(b)

Table 1. Difference limen data from Flanagan's experiment. -DL refers to left shift of formant frequency, and +DL refers to right shift.

(i.e., k=1 in (6) of Appendix C). The resulting two perceptually consistent distance measures may be denoted, for convenience, as d(GM) and d(DC). Scaled distance measures may be similarly denoted as d'(GM) and d'(DC). Among a number of scaling methods that we experimented with, we obtained the best results (in the sense of R1 and R2 being close to unity) with the following definitions:

$$d'(GM) = \frac{\int A(\omega) \left| P_1(\omega) - P_2(\omega) \right| d\omega}{\int A(\omega)\, P_i(\omega)\, d\omega} \quad . \tag{6}$$

$$d'(DC) = \frac{\int L_i(\omega) \left| P_1(\omega) - P_2(\omega) \right| d\omega}{\int P_i(\omega)\, d\omega} \quad . \tag{7}$$

where

$$L_i(\omega) = [P_i(\omega) A(\omega)]^{1/3} . \tag{8}$$

The limits of the integrals in (6) and (7) should be appropriately specified (0 to π for digital spectra, 0 to the highest frequency of interest (we used 5 kHz) for analog spectra). The subscript i in (6) and (7) may be either 1 or 2. For the results that follow, we used i=1. ($P_1(\omega)$ corresponded to the spectrum with formants at their nominal (unperturbed) frequency values.) Table 2 gives R1 and R2 for the two spectral distance measures with and without scaling.

Next, we compared the scaled perceptually consistent measure d'(DC) in (7) with the well-known rms log spectral measure which is defined in Appendix C by (6), where k=2, e(ω) is given by the expression (3) there, and the spectra are GM-normalized. (Notice

Measure	R1	R2
d(GM) (No scaling)	23.13	1.55
d'(GM) (Scaling)	7.55	1.18
d(DC) (No scaling)	21.56	1.46
d'(DC) (Scaling)	3.23	1.09

Table 2. Maximum-to-minimum distance ratio (R1) and AM-to-GM distance ratio (R2) for two distance measures, each considered with and without scaling.

that the rms log measure is not perceptually consistent.) The distance values computed with each of the two measures for the 12 conditions in Table 1 are shown plotted in Fig. 2 after normalizing them by dividing with the mean of the 12 distance values. For either measure, two points are displayed, with a vertical bar connecting them, for each of the 6 formant frequencies; those two points correspond to the left shift (denoted by a small horizontal tic mark) and right shift (denoted by x) of a formant frequency. The midpoints of the vertical bars are connected by a dashed line for d'(DC) and by a solid line for the rms log measure. For a perceptually consistent distance measure, the vertical bars should be short, and the line passing through their midpoints should be nearly flat. Fig. 2, therefore, illustrates the superiority of d'(DC) over the rms log spectral measure.

A different way of viewing the above experimental data is depicted in Fig. 3. Here, we have plotted the spectral distance as a function of the magnitude of the formant frequency shift. Points enclosed by small circles correspond to the rms log measure, and points marked by x, to d'(DC). As expected, encircled points lie (approximately) on a straight line; therefore, the spectral distance in this case depends solely upon the amount of frequency shift, and specifically it does <u>not</u> depend upon the location of the shifted formant relative to other formants. This clearly illustrates the perceptual inconsistency of the employed rms log measure. On the other hand, the points marked by x corresponding to the perceptually consistent measure

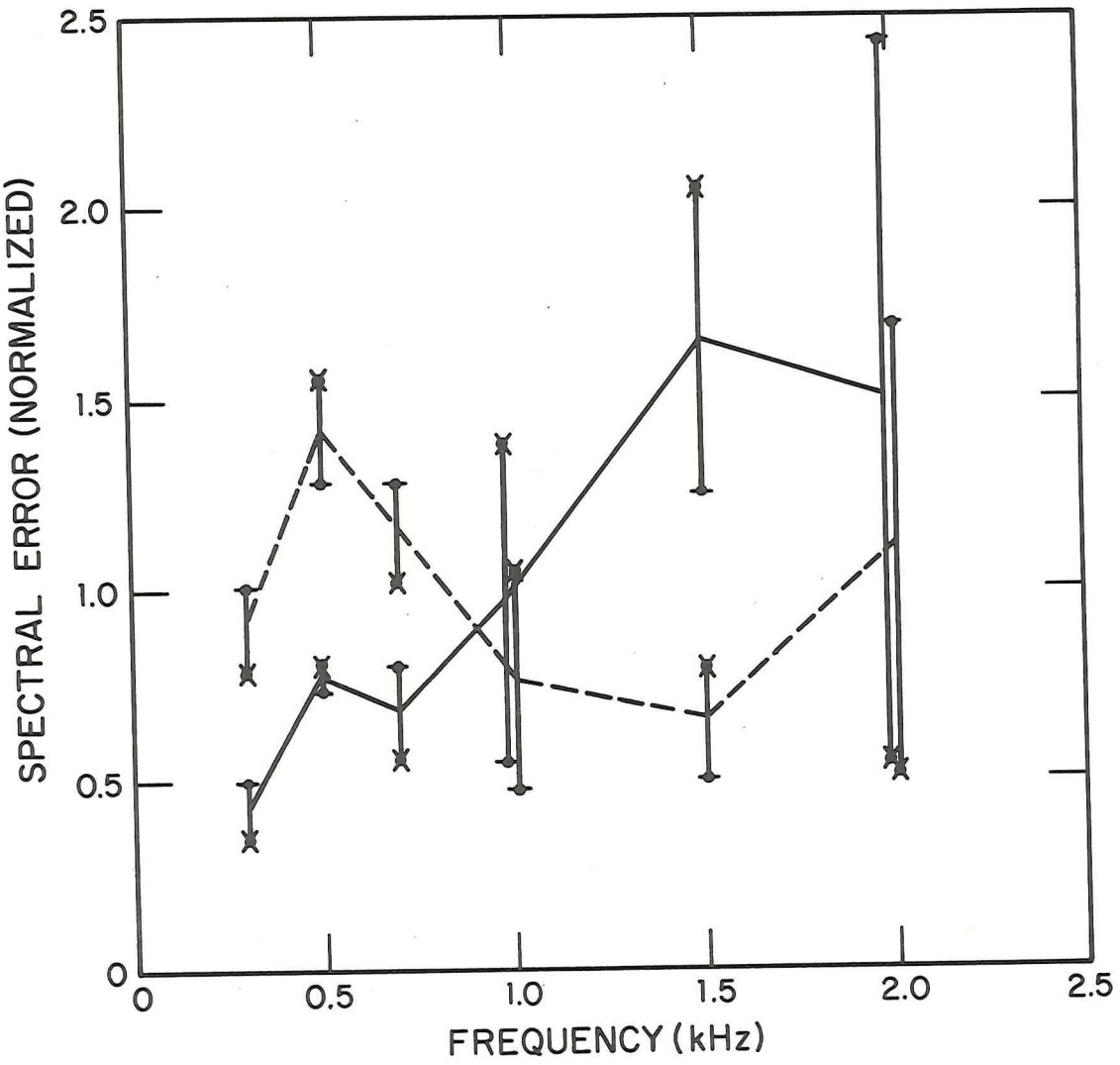


Fig. 2 Plots of spectral deviation due to DL shifts in formant frequency, for two distance measures. (See text for details.)

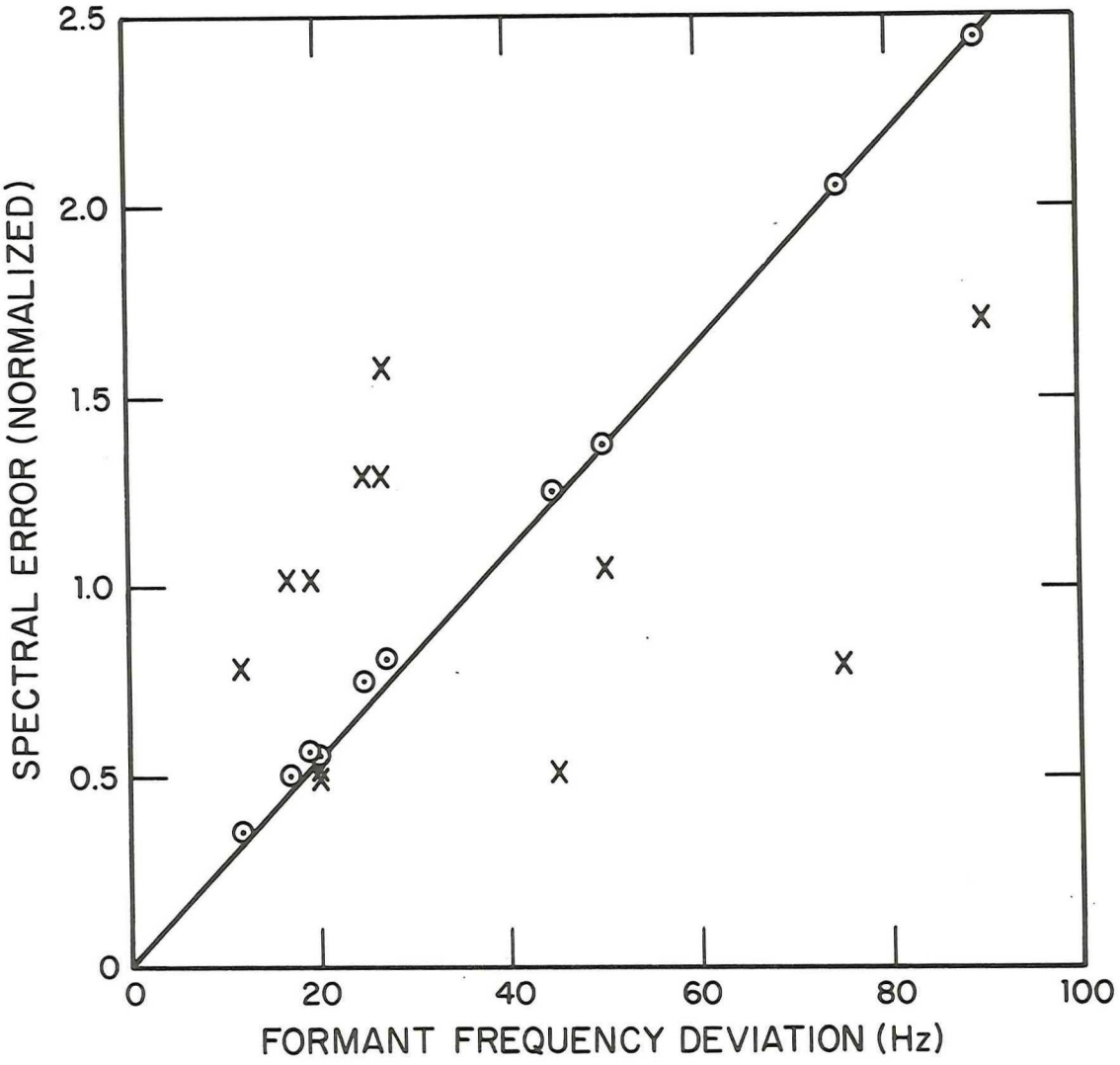


Fig. 3 Plots of spectral deviation due to DL shifts in formant frequency as a function of the magnitude of formant frequency shift, for two distance measures. (See text for details.)

d'(DC) exhibit approximately a flat characteristic.

D. Spectral Distance with Single-Formant Spectra

We studied the characteristics of the two perceptually consistent scaled measures d'(GM) given by (6) and d'(DC) given by (7) as applied to single-formant spectra. This study was motivated by the work of K.N. Stevens ("The Perception of Sounds Shaped by Resonant Circuits", Ph.D. Thesis, MIT, 1952.), which was concerned with the subjective perception of synthetic single-formant stimuli.

Fig. 4 shows the plot of spectral deviation as a function of formant frequency due to a 5% shift in the latter. The formant bandwidth was increased with frequency as shown in Fig. 5; this type of bandwidth variation was observed by Dunn ("Methods of Measuring Vowel Formant Bandwidths", J. Acoust. Soc. Amer., Vol. 33, pp. 1737-1746, Dec. 1961.) Spectral deviation due to 20% shift of the bandwidth of a 1000-Hz formant is shown plotted in Fig. 6 for different bandwidth values.

Finally, we computed the spectral deviations due to DL shifts in formant frequency and bandwidth; the employed DL values were taken from the thesis by Stevens. These are plotted in Fig. 7 (formant frequency shift) and Fig. 8 (formant bandwidth shift).

We have not as yet analyzed the results presented above for single-formant spectra.

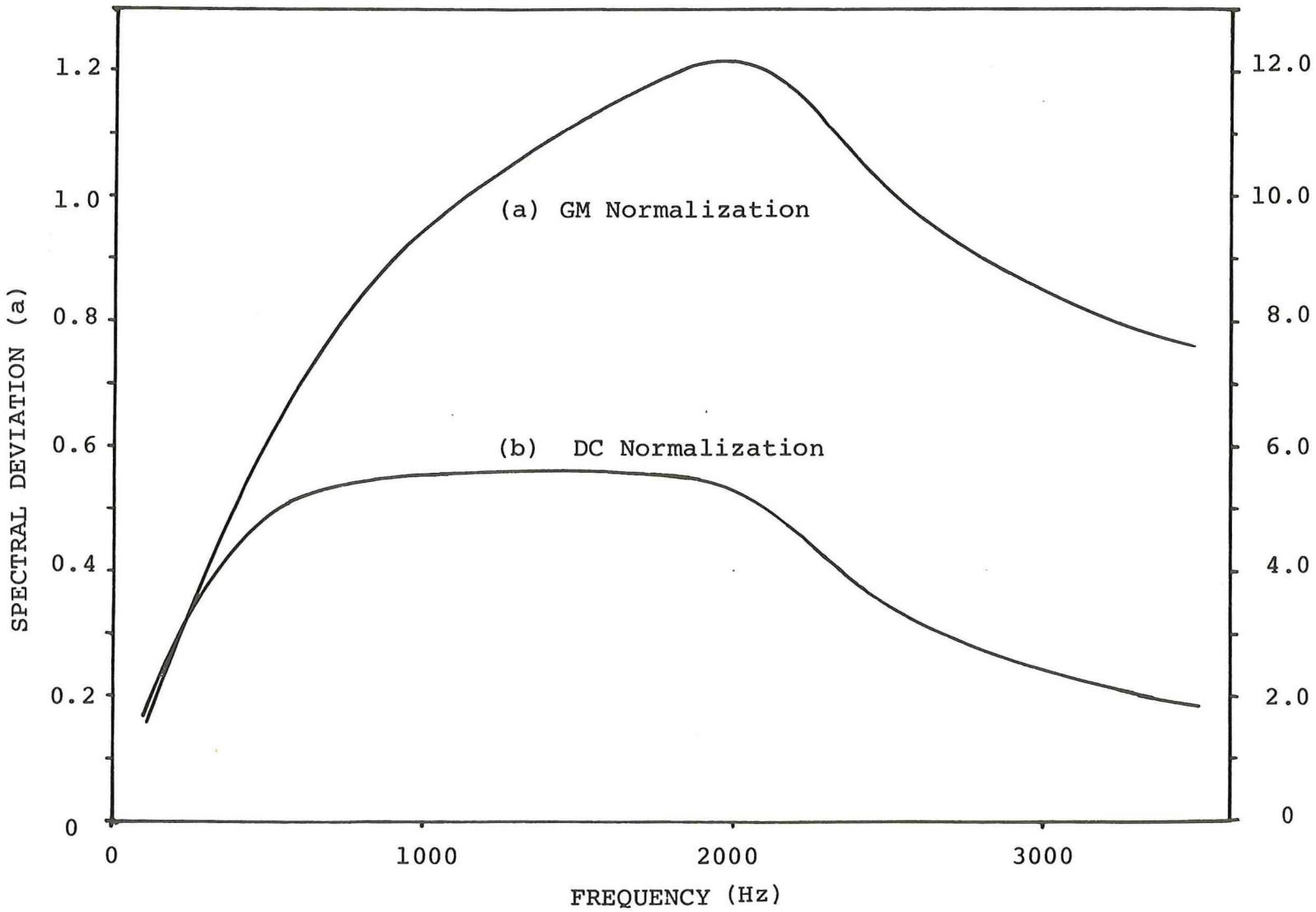


Fig. 4 Spectral deviation due to a 5% shift in frequency of a formant with its bandwidth increasing with frequency

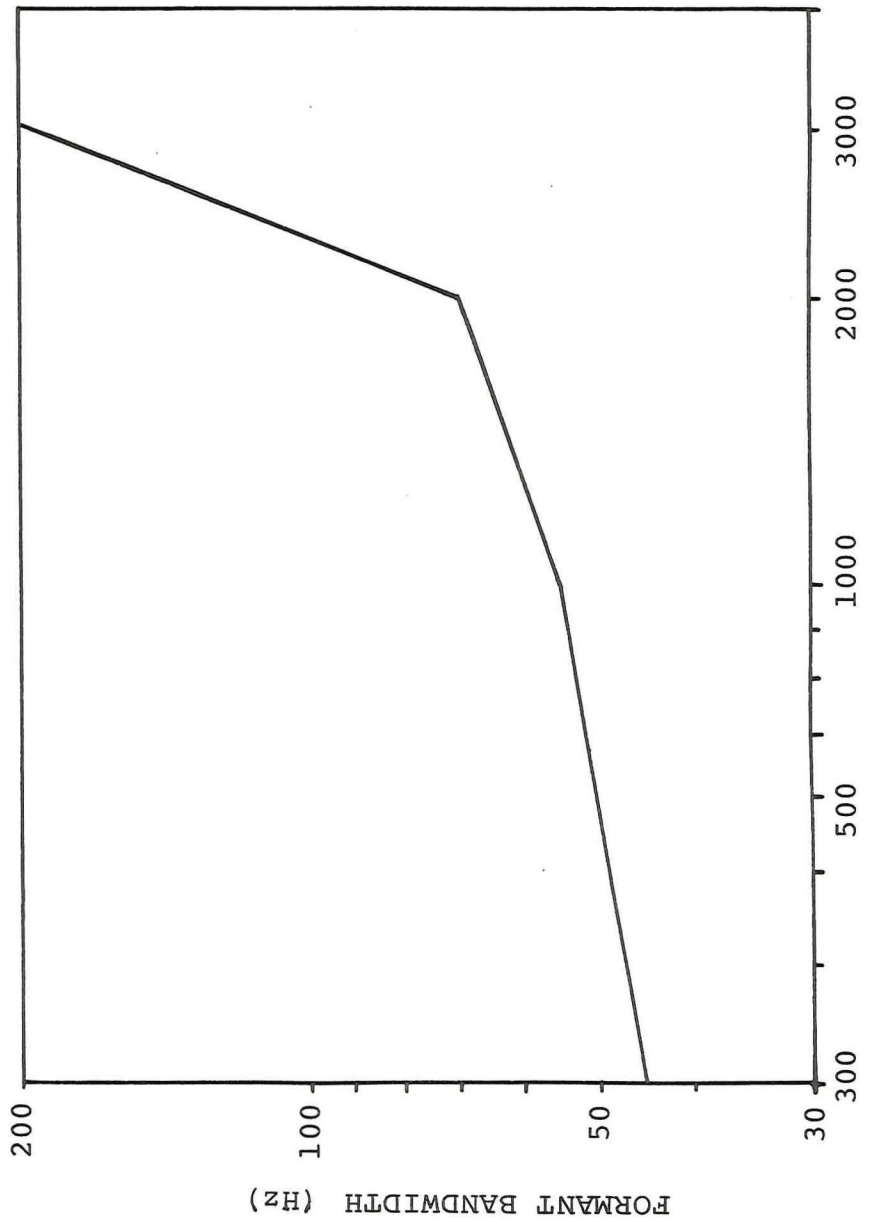


Fig. 5 Approximate variation of formant bandwidth with formant frequency, as suggested by Dunn.

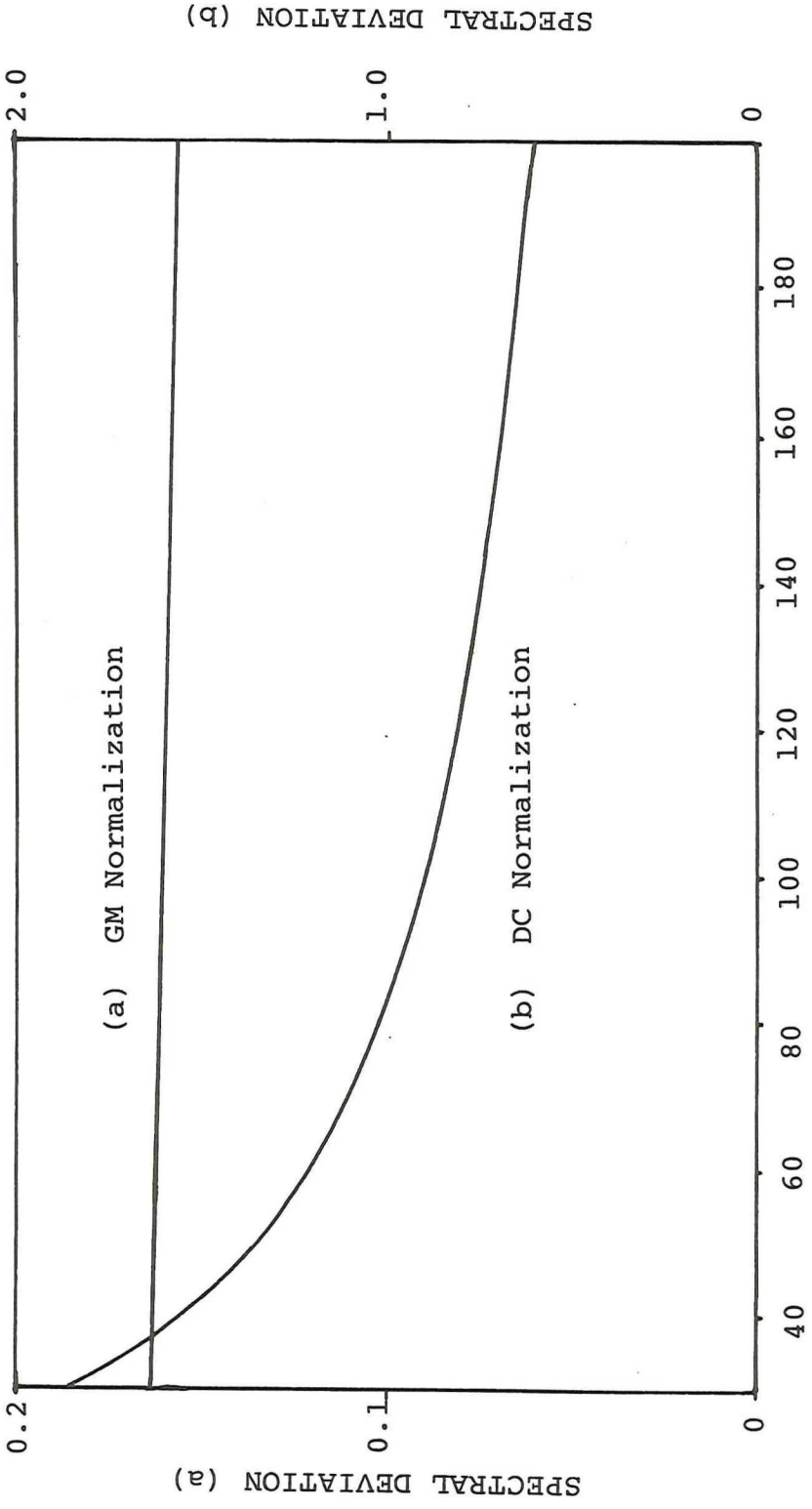


Fig. 6 Spectral deviation due to 20% shift of the bandwidth of a 1000-Hz formant

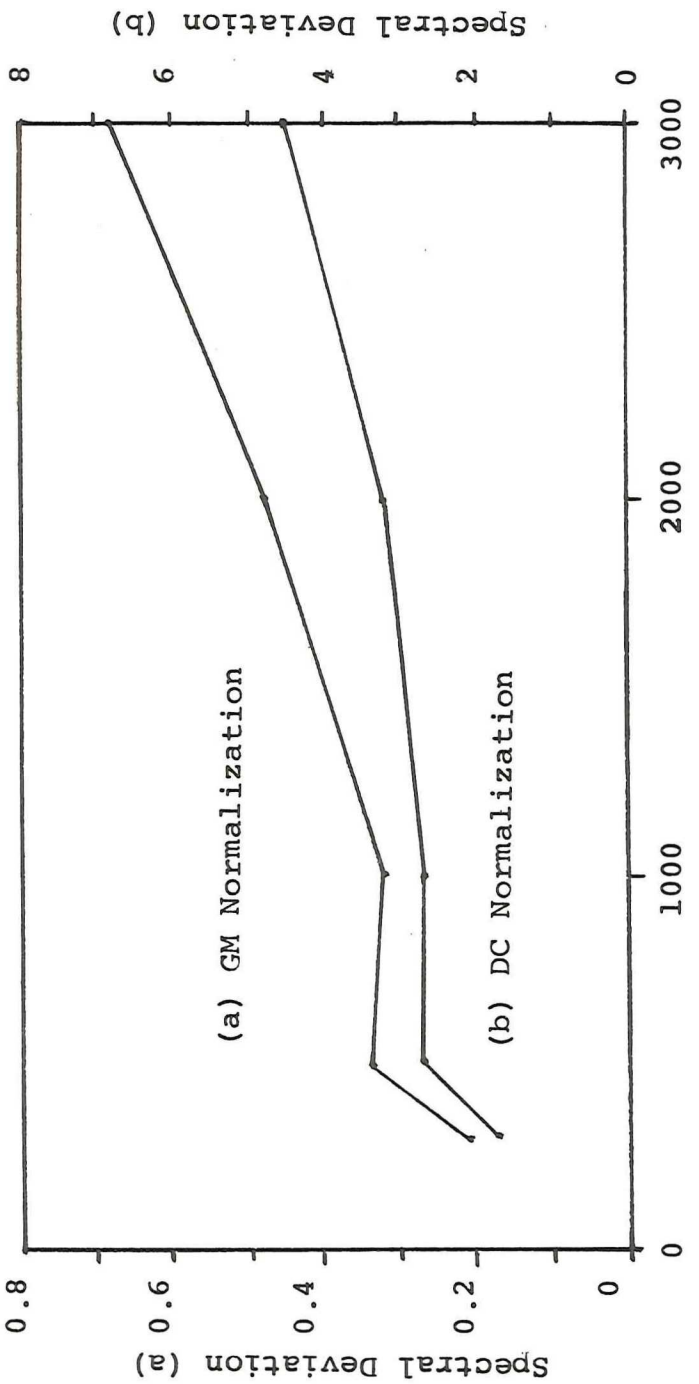


Fig.7 Spectral deviation due to DL shifts in formant frequency with its bandwidth fixed at 48 Hz.

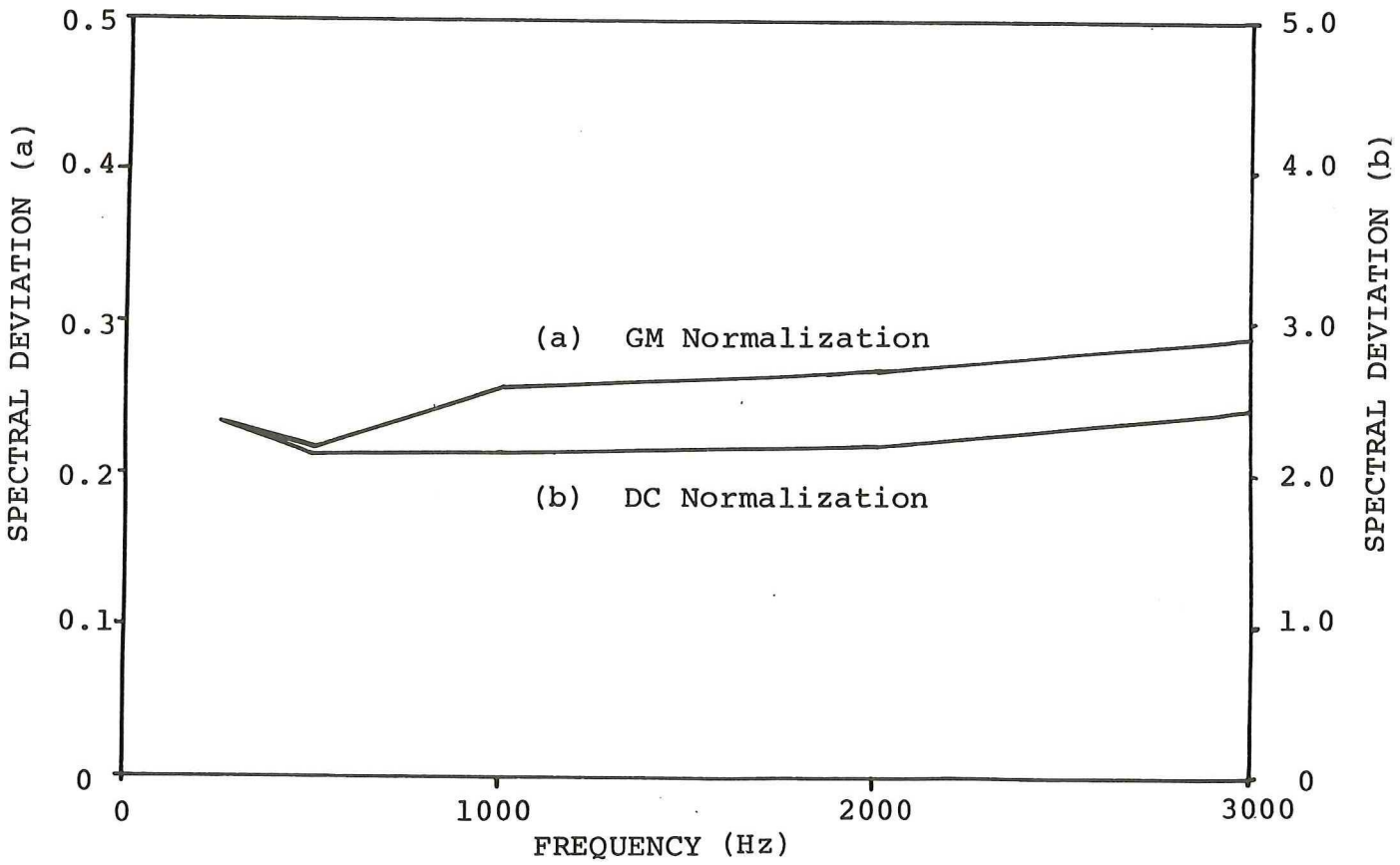


Fig. 8 Spectral deviation due to DL shifts in BW of a formant with its nominal bandwidth at 48 Hz.

V. REAL-TIME IMPLEMENTATION

During the last quarter, we purchased an RT11 operating system for our PDP11/40. Upon delivery of this system, it was modified to permit the use of the existing Telefile/Century Data disc. Work is in progress for writing an FTP (File Transfer Protocol) for the RT11 system to facilitate transfer of digitized signals to and from TENEX. We plan to modify the network speech compression software, developed primarily by ISI, to run with RT11. We plan also to bring up the array processing library supplied by SPS, Inc. and modified by SCRL. The overall system when completed, will be a real-time facility for use in experiments with network LPC algorithms.

VI. SPEECH QUALITY EVALUATION

In the last quarter we have run some pilot studies impinging on the design of our next study of subjective quality. We have now settled on a provisional design of the study, and we are in the process of generating the stimuli for it.

A. Pilot Factorial Study of Variable Frame-Rate Systems.

The variable frame-rate systems used in our last subjective quality study (see QPR #4) did not perform as well as expected: in particular, they failed to perform as well as the fixed rate systems on the slowly varying sentences. The reason for this is fairly obvious, in retrospect. The thresholds for the variable rate systems were set so as to equate the bit rate, averaged over all speaker x sentences, with the bit rates of the fixed rate systems, 2600 bps (see QPR 4, Part III, Appendix Table 1b). This decision was made because the purpose of our first subjective quality tests was to develop the method, rather than pick the optimal parameter values.

Since the <u>average</u> frame rate of the variable rate systems was equal to that of the fixed rate systems, the <u>actual</u> bit rate of the VFR systems was higher than the average on the rapidly changing sentences (e.g. Sentence 4: Which tea-party did <u>Baker</u> go to?), and lower than the average on the slowly varying sentences (e.g. Sentence 1: Why were you away a year, Roy?). This difference in bit rate provides a sufficient explanation for

the relative performance of fixed and variable rate systems, so there was little evidence in our earlier results that the variable rate systems were in fact superior. Our pilot study was performed to check this, before we included the 15 variable rate systems in our main study (described below).

The materials we used had been prepared for an earlier pilot study with a quite different purpose: that of generating an inventory of descriptor terms for the different dimensions of speech quality (see QPR #2, Part III, p.20). The materials consisted of the two "general" sentences (#5: The little blankets lay around on the floor; #6: The trouble with swimming is that you can drown), each spoken by one male (JB) and one female (RS) talker. Each of the four sentences was passed through 8 vocoder systems in a 2x2x2 factorial design. Each system had either 12 or 9 poles (including one used for pre-emphasis), 0.5 dB or 2.0 dB quantization step size, and variable rate threshold of 0 dB or 2.5 dB. The frame rate with 0 dB threshold was 50/sec, and the 2.5 dB threshold produced an average frame rate of about 23/sec. The bit rates ranged between 3154 bps for the best system, to 1060 for the worst (including pitch and gain, but without Huffman coding), and there was a considerable range of quality. Because of the factorial design, these materials were ideal for a pilot comparison of bit saving by variable rate as opposed to varying the number of poles or quantization step size. Details of the systems are given in Table 3.

System	Poles	Quant	VFR* Thrsh	bps	Ratings Mean	(N=128) S.D.
A	12	0.5 dB	0 dB	3157	1.93	1.35
B	12	0.5	2.5	1831	3.22	1.30
C	12	2.0	0	2355	4.03	1.50
D	12	2.0	2.5	1446	3.99	1.44
E	9	0.5	0	2521	3.76	1.91
F	9	0.5	2.5	1456	4.73	1.72
G	9	2.0	0	1771	4.83	1.53
H	9	2.0	2.5	1119	5.20	1.54

*VFR Threshold of Ø dB yields a fixed frame rate of 50/sec (labelled "Fix" in Fig. 9); VFR Threshold of 2.5 dB yields a variable frame rate of 23.3/sec (labelled "VFR" in Fig. 9).

Table 3. System parameters and quality ratings of the eight factorial systems

The thirty-two stimulus sentences were dubbed from Language Master cards onto a stimulus tape, each occurring four times. Presentation order, and sequence effects, were appropriately counterbalanced. Eight subjects made seven-point category ratings of degradation (i.e., large numbers corresponded to more degradation). The results are shown in the two columns at the right of Table 3. In Fig. 9, the mean quality rating is plotted against average bit rate for each of the eight systems. All pairs of systems that differ on only one parameter value are joined by lines. The solid lines join the systems that differ only in the threshold for variable frame rate transmission; the dotted lines join systems that differ only in quantization step size; and the dashed lines join systems that differ only in number of poles. For all pairs of systems joined by lines, the system with the more poles, or the finer quantization, or the faster frame rate, and the higher bit rate, lies to the RIGHT. First, compare the effects on rated quality of the three alternative methods of reducing the bit rate of the best system (the rightmost point in Fig. 9). Changing the variable frame rate threshold from 0 dB (i.e. fixed rate of 50/sec) to 2.5 dB yields a larger saving in bit rate, and a smaller loss of quality, than either coarsening the quantization step size from 0.5 dB to 2.0 dB, or reducing the number of poles from 12 to 9.

The difference in quality was tested for significance, by t-test, for all 12 pairs of systems joined by lines. The values of t, and of F, the variance ratio, are given in Table 4. All

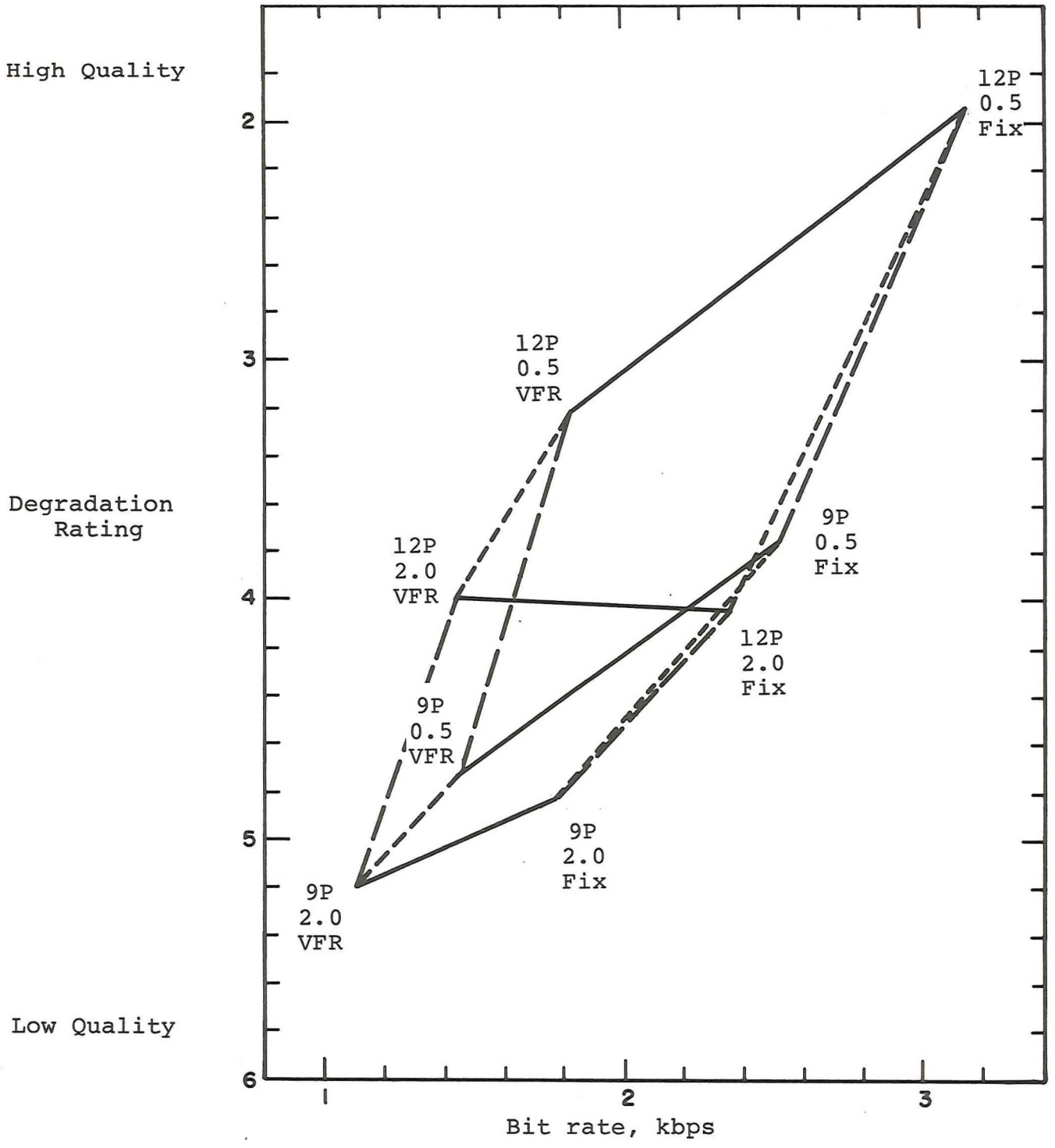


Fig. 9 Mean degradation rating is plotted against average bit rate for the eight systems. The labels by each point give the number of poles (12P or 9P), the quantization step size in dB (0.5 or 2.0). Frame rates are identified as "Fix" (Variable Frame-Rate threshold = 0 dB), and as VFR (VFR threshold = 2.5 dB).

	B	C	D	E	F	G	H
A	7.763*** 1.08	11.77*** 1.23		8.84*** 2.00**			
B			4.51*** 1.22		7.95*** 1.74*		
C			0.21 1.09			4.21*** 1.04	
D							6.46*** 1.14
E					4.30*** 1.24	4.95*** 1.57	
F							2.26* 1.25
G							1.92 1.01

*t > 1.98, p < .05 ***t > 3.37, p < .001

*F > 1.61, p < .05 **F > 1.95, p < .01

Table 4. Values of t (t-test) and F (variance ratio) for the significance of quality differences between pairs of systems.

the differences of quality but two are significant, and all but three are highly significant ($P<.001$). In every case where a comparison can be made, the variable rate systems either yield a greater savings of bit rate, or a smaller degradation of quality, or both, than either of the other two alternatives. Two pairs of systems yielded significant values for F, the variance ratio. These values result from the fact that ratings of systems E and F were bimodal, with relatively little degradation associated with the female speaker, and much more degradation of the male voice. This replicates a result obtained in our first subjective quality study (see QPR #4).

This result clearly supports our previously unproved assertion that variable frame-rate systems can yield substantial savings of bit rate, with little loss of quality.

B. Design of a New Subjective Quality Study.

The main motivations for the new study were 1) to provide baseline data for a new series of objective quality measures; 2) to study the effects on quality of trading off the number of poles, the quantization step size, and the frame rate against each other, in such a way that we could state which combination would yield the best quality for a range of overall bit rates; and 3) to document our claim that variable frame rate transmission can yield substantially lower bit rates without appreciable degradation of quality.

The main purpose of our earlier tests of quality was to develop the method, rather than decide on the best combination of parameters for high quality. As a result, the study was not a factorial design, and the range of quality in the 13 systems was too small to provide the level of reliability we need, in our subjective data, for the development and calibration of objective quality measures. This was largely a result of our decision to equate the bit rates of all systems at 2600 bps. Furthermore, to achieve this aim, the three variables we varied (no. of poles, frame rate, and quantization step size) were always traded off against each other. That is, there were no pairs of systems that differ on only one of the three dimensions. In view of these facts, it is surprising that our tests were as successful as they were in separating out the second and third perceptual dimensions (see QPR #4). The separation might well have been more dramatic, and might in addition have produced a fourth dimension (corresponding to quantization step size), had we abandoned our commitment to 2600 bps, and varied one dimension at a time.

Our initial design consisted of 64 fixed frame-rate vocoder systems, in a 4x4x4 factorial design, plus 18 variable frame-rate systems. All 82 systems were to use an all-pass filtered pulse/noise signal as the excitation for the synthesizer. Unfortunately, this design confounds the effects on quality of varying the encoder parameters, with the effects caused by the use of the pulse/noise excitation model. We therefore ran a pilot experiment, to decide whether we should use the residual

error signal as the excitation source. We processed several sentences through the best and worst of the proposed 64 fixed rate systems, and compared the effects on quality of using the unquantized residual error from the best system as the excitation source, with using the pulse/noise excitation. The results of the pilot experiment showed that the range of quality between the best and worst systems, using residual excitation, was too small to yield a meaningful separation of 82 systems without an inordinate number of replications. A further disadvantage of the residual excited systems is that, although the bit rate is smaller than that of the original speech, it is not greatly smaller, due to the high bit rate of the residual error signal. Therefore, the bit rates of the 82 systems would be bunched together, and a choice of the optimum set of encoder parameters as a function of bit rate could not be made.

We rejected the solution of measuring the quality of all 82 systems with <u>both</u> types of excitation, because the study would have become unmanageably large. Instead, we eliminated 15 of the 64 fixed-rate pitch-and-gain excited systems, leaving 49 (4x4x3+1), and added in their place a subset of 16 of the 49, excited by the unquantized residual error signal. The parameter values were as follows:

Rank	1	2	3	4
Number of Poles	13	11	9	8
Quantization Step Size, dB	0.25	0.5	1.0	2.0
Frame rate, per second	100	67	50	33

Note: The number of poles quoted is the number of transmitted coefficients, and excludes the single pole used for preemphasis.

The particular choice of parameter values given above leads to a uniform 1-bit separation for the variable b, the number of bits per (voiced) frame per coefficient. For example, with 11 poles, the above step sizes yield b values of 6, 5, 4, and 3. Table 5 lists values of b and B (bits per voiced frame), for various combinations of number of poles (p) and quantization step size.

The parameters of the 80 selected systems are given in Table 6. Their bit rates (excluding residual excited systems) range from 8700 bps, for the best fixed rate system, down to 1254 bps for the worst, and from 3600 bps for the best variable rate system, down to 1750 for the worst. These figures include pitch and gain, but not the benefits of Huffman coding, which could reduce the quoted rates by about 20%.

Seven sentence tokens will be used, as follows: JB1, DD2, RS3, AR4, JB5, DK6, RS6. The sentences were chosen from the full set of 36 under the following constraints: a) the corresponding vectors should be as widely spread as possible in the results of our earlier multi-dimensional study (see QPR #4), to ensure a wide range of both phonetic and speaker variation , b) each sentence should occur at least once, c) each speaker (except PF, our slowest speaker) should be represented at least once, d) four sentences should be spoken by males, three by females, and

		poles			
		13	11	9	8
step size	.25 dB	5.846 (76)			
	.5	4.846 (63)	5 (55)	5.222 (47)	5.375 (43)
	1	3.846 (50)	4 (44)	4.222 (38)	4.375 (35)
	2	2.846 (37)	3 (33)	3.222 (29)	3.375 (27)

Table 5. Bits per (voiced) frame per coefficient for various values of number of poles (p) and quantization step size. (The figures in one row increase as p decreases, because more bits are used for low-order than high-order coefficients). Numbers enclosed in parentheses give bits/frame, excluding pitch and gain.

No. of Systems	Excit	No. Poles	Quant, dB	Frm/sec	VR Thrsh
1	Pitch & Gain	13	0.25	100	
48	Pitch & Gain	13,11,9,8	0.5,1.0,2.0	100,67,50,33	
1	Residual	13	0.25	100	
6	Residual	13,9	1.0	100,50,33	
9	Residual	11	0.5,1.0,2.0	100,50,33	
6	VR, P & G	13,9	1.0		1.0,1.5,2.0
9	VR, P & G	11	0.5,1.0,2.0		1.0,1.5,2.0

Table 6. Provisional system parameters for new subjective tests

e) some sentences with abnormal inflection should be excluded (e.g. JB2).

The experimental design requires some 12-15 subjects to make 6-8 judgments on each of the 560 stimuli. Judgments will be seven or nine-point category scales, and we will use successive intervals scaling to yield unidimensional quality ratings for use in subsequent objective tests. We will also analyze the ratings with MDPREF, to see if the factorial relationships between the system parameters are reflected in the ratings.

APPENDIX A

LPCW: AN LPC VOCODER WITH LINEAR PREDICTIVE SPECTRAL WARPING

NSC Note 89, March 8, 1976

(Authors: John Makhoul and Lynn Cosell)

(This paper was presented at the 1976 International Conference on Acoustics, Speech and Signal Processing, Philadelphia, April 12-14, 1976).

LPCW: AN LPC VOCODER WITH LINEAR PREDICTIVE SPECTRAL WARPING

In ordinary linear prediction the speech spectral envelope is modeled by an all-pole spectrum. The error criterion employed guarantees a uniform fit across the whole frequency range. However, we know from speech perception studies that low frequencies are more important than high frequencies for perception. Therefore, a minimally redundant model would strive to achieve a uniform perceptual fit across the spectrum, which means that it should be able to represent low frequencies more accurately than high frequencies. This is achieved in the LPCW vocoder: an LPC vocoder employing our recently developed method of linear predictive warping (LPW). The result is improved speech quality for the same bit rate.

1. Introduction

Narrow-band LPC vocoders with transmission rates less than 4800 bps have generally dealt with speech sampled at less than 10 kHz and usually closer to 6.5 kHz. Since the bit rate needed for transmission is roughly proportional to the sampling rate, it is argued justifiably that the possible increase in speech intelligibility and quality in going to 10 kHz is not commensurate with the increase in bit rate, and so sampling rates closer to 6.5 kHz have dominated

the vocoder scene. The argument can also be phrased another way. If the bit rate is to remain fixed (e.g., 2400 bps), then an increasing the number of bits for each frame means that one is forced to transmit fewer frames per second. Thus, while spectral fidelity is increased for each transmitted frame, the accuracy in following the dynamic aspects of the signal is decreased.

Traditional channel vocoder systems have solved this problem by positioning their filters nonlinearly such that more filters are at low frequencies than at high frequencies [1]. It is not unusual to see a filter placed as high as 7 kHz in a channel vocoder. Thus, the total speech bandwidth represented can be about 7 kHz, which is to be contrasted with bandwidths closer to 3 kHz in LPC vocoders. (It should not be concluded from this, though, that channel vocoders produce higher quality speech than LPC vocoders for a given bit rate.)

A hybrid solution was introduced in the TRIVOC vocoder [2], which used an LPC representation at low frequencies and a channel vocoder at higher frequencies. This, of course, has the disadvantage of having to program two different vocoder systems.

This paper presents LPCW: an LPC vocoder that is capable of representing low frequencies better than high frequencies. This suggests the possibility of wide-band

speech at low bit rates.

2. Linear Predictive Warping

The idea behind LPCW is quite simple: Warp the spectrum such that high frequencies are compressed relative to low frequencies, then apply spectral linear prediction [3] to the warped spectrum. Because the resulting representation is uniform across the warped spectrum, it means that low frequencies are better matched than higher frequencies since the latter are compressed.

The procedure for computing the coefficients of the warped spectrum is as follows:

(a) Window the signal and compute its spectrum.
(b) Warp the spectrum as desired.
(c) Take the Fourier transform of the warped spectrum to get the autocorrelation R(i).
(d) Solve for the predictor parameters from the normal equations:

$$\sum_{k=1}^{p} a(k)\, R(i-k) = -R(i), \quad 1 \le i \le p, \qquad (1)$$

where a(k) are the predictor coefficients and p is their number. The reflection coefficients, which are obtained as a byproduct of the solution, can be converted to log area ratios, then quantized and transmitted [4].

In warping the spectrum, it is practical (because of FFT algorithms) to compute the spectral values at equally spaced frequencies. This can be done by simple interpolation from the signal spectral values. The

autocorrelation R(i) can then be computed via the FFT.

The procedure given above for linear predictive warping makes use of the autocorrelation method of linear prediction [5]. If the analysis is done using the covariance, lattice, or covariance lattice [6] methods, then the procedure has to be modified as follows: after solving for the predictor coefficients, compute the all-pole model spectrum, then continue the procedure starting at step (b) above. The all-pole model spectrum is given by:

$$P(\omega) = \frac{1}{\left|1+\sum_{k=1}^{p} a(k)e^{-jk\omega}\right|^2} . \qquad (2)$$

3. Spectral Dewarping

At the receiver of the LPCW vocoder, the received parameters are decoded. If log area ratios are received, they are decoded into reflection coefficients, which are converted in turn to the corresponding predictor coefficients using a simple recursive procedure [5]. These coefficients correspond to the warped spectrum and, therefore, cannot be used for synthesis. One must first perform the necessary dewarping.

The dewarping procedure is as follows:

(a) Using the decoded predictor coefficients, compute the all-pole model spectrum from (2).
(b) Dewarp this spectrum using the inverse of the function

used in the original warping.
(c) Take the Fourier transform of the dewarped spectrum to obtain the corresponding autocorrelation function.
(d) Use this autocorrelation function in (1) to compute the predictor coefficients (and hence the reflection coefficients) corresponding to the dewarped spectrum. The number of poles (predictor coefficents) here can be as large as desired to approximate the dewarped spectrum.
(e) Synthesize the speech waveform using these computed coefficients.

After step (b) above it is possible to take a different route to obtain the parameters of the synthesis filter. Instead of using linear prediction, one could use the cepstrum [7] to compute the minimum phase impulse response whose spectrum is identical to the dewarped spectrum. This impulse response is then used for the synthesis filter.

Discussion

It is clear from the dewarping procedures given above that the amount of processing needed at the synthesizer is comparable to that of the analysis. This increase in computation relative to a regular LPC vocoder is certainly a disadvantage. Whether the extra expense is justified or not depends on the benefits achieved. For a given bit rate, the main benefit is an increase in the speech bandwidth representable using the same bit rate. This increase in bandwidth is on the order of 50%.

4. Warping Functions

Since linear predictive (LP) warping allows for arbitrary warping of the spectrum, one must choose a warping function appropriate for vocoder purposes. One reasonable warping function would transform the linear frequency scale to the mel scale [8], which compresses high frequencies relative to low frequencies.

The relation between the mel scale and frequency is shown in Fig. 1, which shows how subjective pitch (in mels) is related to frequency (in Hz) for pure tones up to 5 kHz. This relation is similar to those of critical band masking effects and equal intelligibility curves [8]. The mel-frequency relation can be approximated by the following equation

$$m = 2595 \log_{10}(1+\frac{f}{700}), \qquad (3)$$

where f is the frequency in Hz and m is the pitch in mels. The mel scale is adjusted such that m=1000 mels corresponds to f=1000 Hz.

Since, in our application, spectra are defined in the z plane, we need a warping function on the angle (which corresponds to frequency) in the z plane. This implies that we must assume a particular sampling frequency F. Let

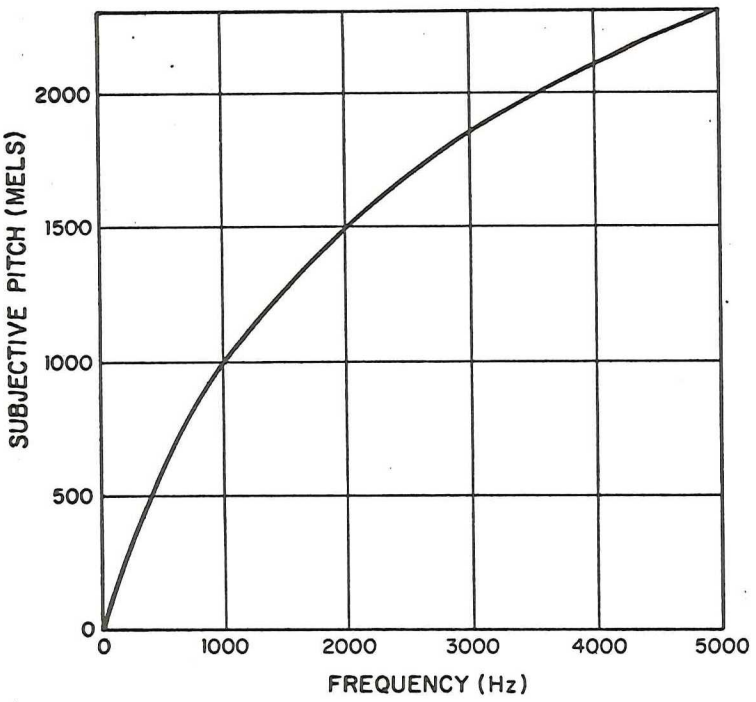


Fig. 1. Subjective pitch versus frequency.

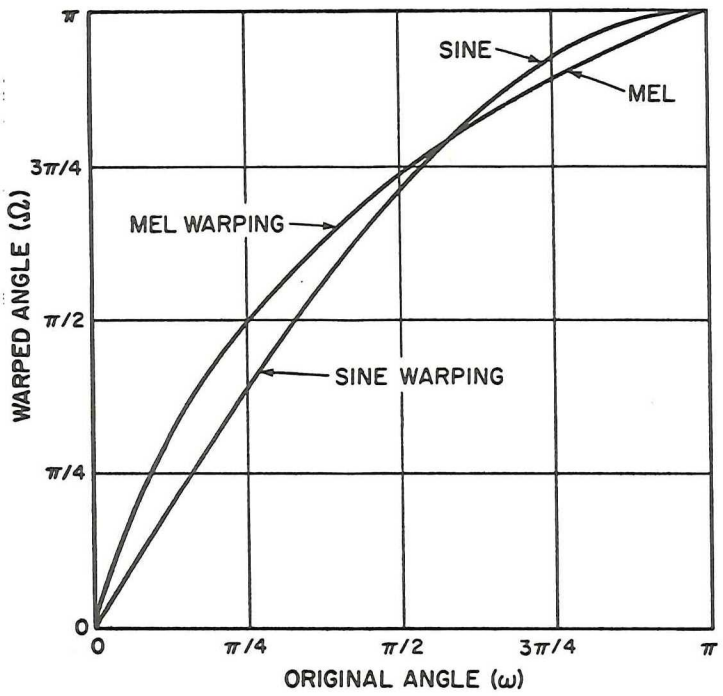


Fig. 2. Mel warping and sine warping.

$\omega = 2\pi\frac{f}{F}$ = original angle in the z plane corresponding to frequency f,

Ω = warped angle corresponding to f.

The warping function Ω is obtained from (3) by setting $\Omega=\pi$ for $\omega=\pi$ or f=F/2, half the sampling frequency. The result is:

$$\Omega = \pi \frac{\log_{10}(1+\frac{f}{700})}{\log_{10}(1+\frac{F}{1400})}, \quad 0\leq f\leq\frac{F}{2} . \tag{4}$$

Note that the warping function in (4) is defined only up to f=F/2. For $F/2\leq f\leq F$, the function is taken to be the mirror image about the real axis. The mel warping function is plotted in Fig. 2 for a sampling frequency F=10 kHz, which corresponds to a speech bandwidth of 5 kHz.

The mel warping function could be used very profitably with a homomorphic vocoder [9] which employs cepstral warping or autocorrelation warping [10]. However, using the mel function with an LPCW vocoder seems to give unsatisfactory results. We believe the reason to be as follows. For LP to give best results, it is important that the all-pole model is well suited to the signal spectrum, which is true for a large and perceptually important class of speech spectra. If the signal spectrum is warped nonlinearly, then the all-pole model ceases to be a good spectral model. Therefore, the results are bound to be less than satisfactory. Note that this problem does not affect

cepstral warping results, since cepstral warping is not based on a specific model.

The solution we offer to this problem in an LPCW vocoder is to have a warping function that is as linear as possible in the frequency range where the all-pole model is important, e.g. up to the third formant region. For higher frequencies the function can be quite nonlinear since only a rough estimate of the spectrum at those frequencies is needed. Fig. 2 shows a sine warping function

$$\Omega = \pi \sin(\frac{\pi f}{F}), \quad 0 \leq f \leq \frac{F}{2}, \tag{5}$$

which, for F=10 kHz, is nearly linear up to 2.5 kHz, and very nonlinear above that. One could, of course, design other warping functions that comprise more than a single curve.

5. Examples

Figs. 3 and 4 show two examples of using the sine warping function with spectra of the vowel [o] and the fricative [s], respectively. In each of the two examples, Fig. a is a 12-pole fit to the original spectrum, Fig. b is a 9-pole fit to the warped spectrum (shown after dewarping), and Fig. c is a 9-pole fit to the original spectrum. Note the greater detail in the first formant region in Fig. 3b as compared to Fig. 3a, while the high frequency region is not

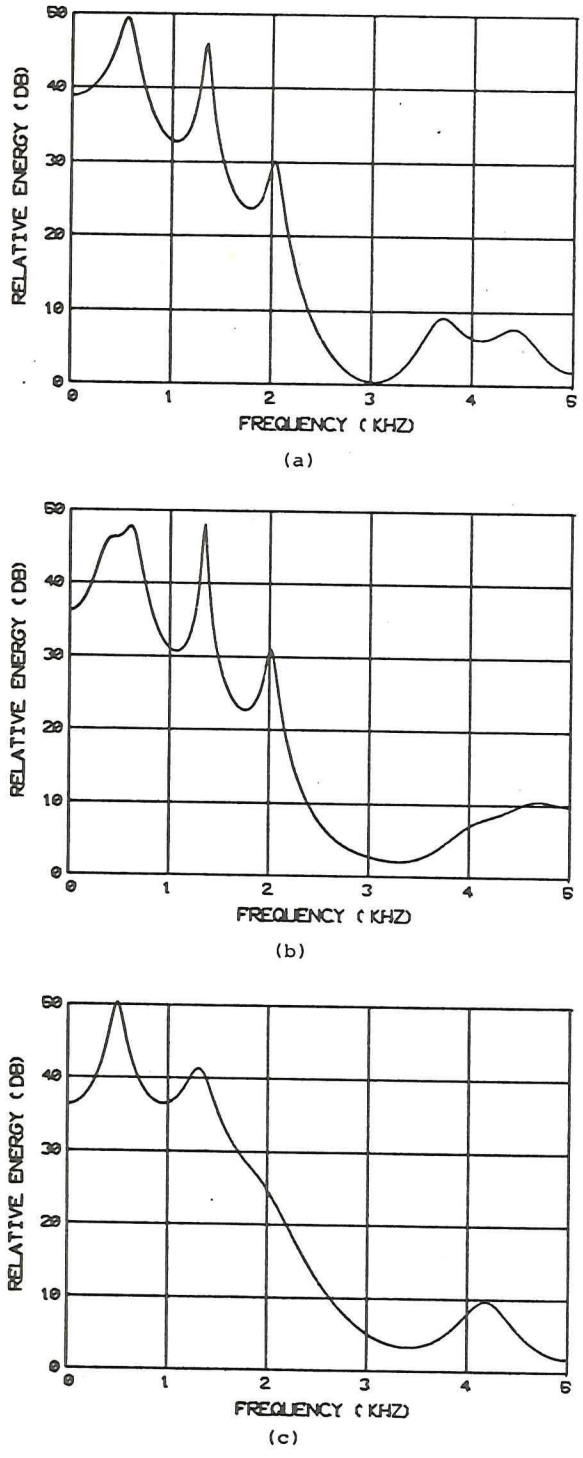


Fig. 3. LP spectra for the vowel [o].
a) Original, 12-pole
b) Warped, 9-pole (shown dewarped)
c) Original, 9-pole.

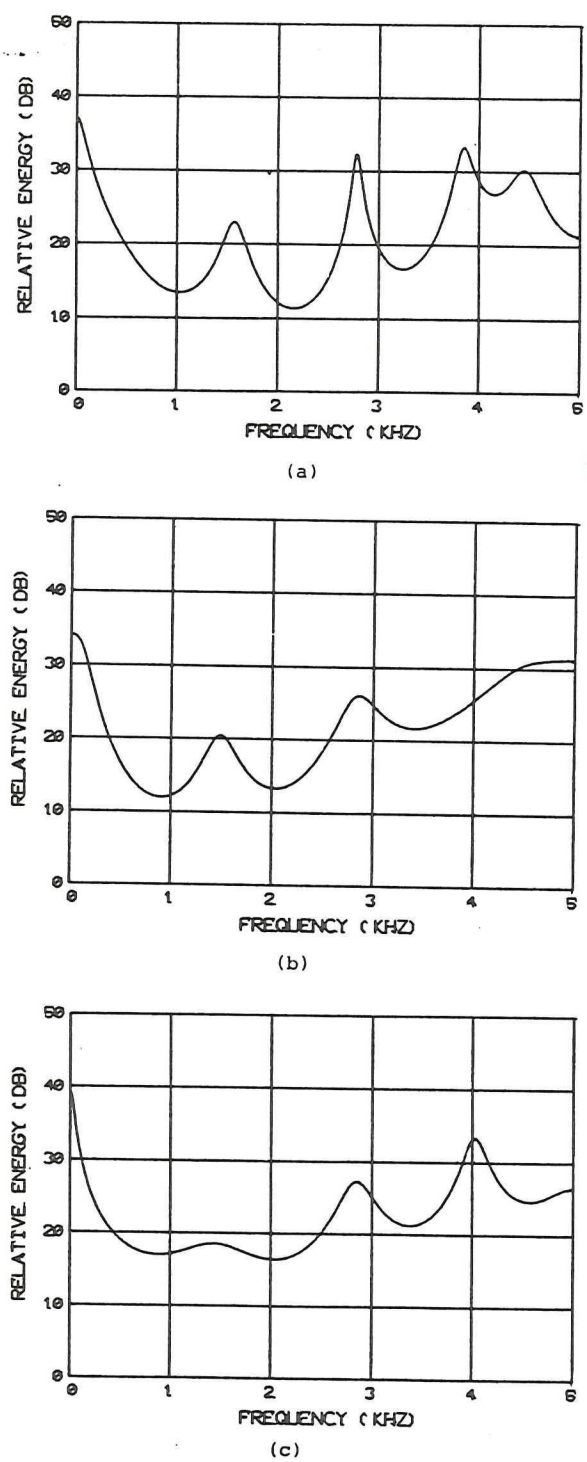


Fig. 4. LP spectra for the fricative [s].
a) Original, 12-pole
b) Warped, 9-pole (shown dewarped)
c) Original, 9-pole.

matched as well in Fig. 3b. In comparing the two 9-pole fits, Figs. 3b and 3c, there is no doubt that Fig. 3b is a better "perceptual" fit to the spectrum, since the first three formants in Fig. 3b are better matched than in Fig. 3c. In contrast, Fig. 4c seems to be a better fit to the spectrum than 4b. But for a fricative, the match of Fig. 4b might be enough for good quality resynthesis.

These examples demonstrate that the use of spectral warping with an LPC vocoder could lead to a more efficient representation of the spectrum for the same speech quality. Although it might be practical to employ a fixed warping function for all situations, it is certainly possible to use several warping functions for different types of spectra. However, it is not clear that the possible increase in efficiency is worth the extra cost.

6. Conclusions

LPCW, an LPC vocoder with LP spectral warping, has been proposed. In this vocoder, a spectral warping function is used to compress high frequencies relative to low frequencies; a technique which is hypothesized to accommodate wider band speech signals. The result is improved speech quality for the same transmission rates.

References

1. J.L. Flanagan, Speech Analysis Synthesis and Perception, Second Edition, New York: Springer-Verlag, 1972.

2. J. Roberts, C. Smith and R. Wiggins, "Triple-Function Voice Coder," J. Acoust. Soc. Am., Vol. 57, Supplement No. 1, S35, Spring 1975.

3. J. Makhoul, "Spectral Linear Prediction: Properties and Applications," IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-23, 283-296, June 1975.

4. R. Viswanathan and J. Makhoul, "Quantization Properties of Transmission Parameters in Linear Predictive Systems," IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-23, 309-321, June 1975.

5. J. Makhoul, "Linear Prediction: A Tutorial Review," Proc. IEEE, Vol. 63, 561-580, April 1975.

6. J. Makhoul, "New Lattice Methods for Linear Prediction," IEEE Int. Conf. Acoustics, Speech and Signal Processing, Philadelphia, Pa., April 1976.

7. A. Oppenheim and R. Schafer, Digital Signal Processing, Ch. 10, New Jersey: Prentice-Hall, 1975.

8. J.C.R. Licklider, "Basic Correlates of the Auditory Stimulus," in Handbook of Experimental Psychology, (S.S. Stevens, ed.), 985-1039, New York: John Wiley and Sons, 1951.

9. A. Oppenheim, "A Speech Analysis-Synthesis System Based on Homomorphic Filtering," J. Acoust. Soc. Am., Vol. 45, 458-465, Feb. 1969.

10. J. Makhoul, "Methods for Nonlinear Spectral Distortion of Speech Signals," IEEE Int. Conf. Acoustics, Speech and Signal Processing, Philadelphia, Pa., April 1976.

APPENDIX B

A FRAMEWORK FOR THE OBJECTIVE EVALUATION OF VOCODER SPEECH QUALITY

NSC Note 86, March 8, 1976

(Authors: John Makhoul, R. Viswanathan and William Russell)

(This paper was presented at the 1976 International Conference on Acoustics, Speech and Signal Processing, Philadelphia, April 12-14, 1976)

A FRAMEWORK FOR THE OBJECTIVE EVALUATION OF VOCODER SPEECH QUALITY

While there exist methods in the literature for objectively evaluating the intelligibility of speech in the presence of stationary noise, little has been done regarding the objective evaluation of either the intelligibility or the quality of vocoded speech. We present a framework within which we have begun a step-by-step program to develop objective measures for vocoded speech quality that are consistent with results from subjective tests.

1. Introduction

The ultimate criterion for determining the quality of the speech that is produced by any compression, encoding or transmission system is the way it sounds to the human listener. Although there are well established procedures to test the intelligibility of speech, little work has been done in developing procedures to test speech quality, and in particular vocoder speech quality. The few procedures that are available are subjective and require extensive testing with human listeners, which is expensive in terms of both time and money.

It would be desirable to develop objective procedures for speech quality evaluation that correlate well with the

scores obtained from subjective listening tests. These objective measures would ensure uniformity in evaluation as well as enable the evaluation to be done by computer. Also, the measures can be used in the design of better quality vocoders. While there exist methods in the literature for objectively evaluating the intelligibility of speech in the presence of stationary noise [1,2], little has been done regarding the objective evaluation of either the intelligibility or the quality of vocoded speech. The problem is that if one regards the distortion in the vocoded speech signal as noise superimposed on the signal, then this noise is not only nonstationary but is correlated with the signal. This makes the problem of objective evaluation of vocoded speech quality a difficult one. However, given the immense long-term benefits in terms of time and expense, any headway into the solution of the problem is desirable.

This paper presents a framework within which we have begun a step-by-step program to develop objective measures of vocoded speech quality that are consistent with results from subjective tests.

2. Necessary Conditions

Let s(n) be the original speech signal and s'(n) a vocoded version of the same signal. Our aim is to develop measures that compare the quality of s'(n) relative to s(n).

Note that the formulation of this problem is different from that of the objective evaluation of speech intelligibility in the presence of noise. In the latter, the noise spectrum is assumed stationary and can be measured directly. The resulting objective intelligibility scores are obtained by comparing the average signal spectrum to the noise spectrum [1,2]. The same procedure cannot be applied in the case of vocoded speech because the "noise" that corrupts the signal is not well defined, and in any case not easily measured. Even if the latter were possible, such noise cannot be considered stationary and it is also correlated with the signal. Therefore, one must somehow compare the vocoded signal s'(n) to the original signal s(n).

One of the main problems in comparing s'(n) to s(n) is that of time synchronization, so that corresponding segments of the two signals can be compared. However, assuming that somehow one is able to align the two signals, the problem of comparing s'(n) to s(n) remains.

In many communication systems, the average mean squared difference error between two signals is taken as a measure of distance or deviation between the two signals. It is simple to show that such an error measure cannot be a measure of the difference in quality between the two signals. This is done by offering a counterexample. Let s(n) be the input to an all-pass filter, and let s'(n) be

its output. The filter can be designed such that the wave shape of s'(n) is quite different from s(n), and such that the mean squared difference between s'(n) and s(n) is large. However, we know from perceptual experiments that, in all likelihood, the difference between s'(n) and s(n) is insignificant as judged by a human listener (at least for vocoder purposes). In fact, it is well known that, except for pitch, phase information is quite irrelevant to the perception of speech [3]. It is difficult to imagine an error criterion on the waveform which would be insensitive to phase.

The answer is clearly to go to the spectrum. In fact, vocoders have traditionally transmitted parameters related to the magnitude of the spectrum. Channel vocoders have used one type of phase realization for synthesis, and LPC vocoders have used another (minimum phase). The problem, then, seems to reduce to a comparison between the short-time spectra of s'(n) and s(n). But the spectrum is only one aspect of the signal that is important to perception and is distorted by the vocoder. The other important aspect is the source information.

After some thought it became clear to us that objective measures for the evaluation of vocoded speech quality must obey two maxims:

(1) They must be a function of the vocoding process, and in particular the vocoder transmission parameters,

(2) They must somehow relate to perception.

The first maxim basically says that the objective evaluation of vocoded speech quality cannot be done abstractly, treating s'(n) as some arbitrary distortion of s(n), but rather it must relate directly to the vocoding process. The second maxim merely states the obvious necessity to have the objective measures be perceptually meaningful. These two maxims not only form a sound basis on which to build these measures, but also offer the hope of a diagnostic tool for the evaluation and refinement of vocoder design. Based on the two maxims, therefore, we proceeded to develop the general framework for objective quality evaluation.

3. Determiners of Quality

In a vocoder system, there are four major identifiable components that can contribute to the degradation of vocoded speech quality: analysis, encoding, transmission, and synthesis. We shall discuss the types of errors introduced by the different components, in an effort to identify the major determiners of speech quality in the vocoding process. This would then give us a handle with which to design our objective measures.

Transmission

Channel transmission errors are an important factor in the choice of a vocoder system, in that different vocoders are affected differently by different types of channel errors. However, given that error correcting codes can reduce sharply the effective error rate, one must still explain the degradation in quality due to the vocoder itself. Therefore, in attempting to develop objective quality measures, we shall assume that channel transmission errors are negligible.

Analysis

The importance of the analysis component is apparent when we consider that it embodies the particular speech model employed. The parameters extracted in this component determine the upper bound on the quality of the synthesized speech.

The general vocoder speech model is that of a source exciting a system that represents the short-time spectrum. We shall restrict our discussion here to LPC vocoders, with the knowledge that it can be extended easily to other types of vocoders (e.g. channel vocoders). The LPC model is that of a source with a relatively flat spectral envelope, exciting an all-pole filter. There are three main types of LPC vocoders, depending on the type of source excitation:

residual excited, voice excited, and pitch excited. However, all three types of vocoders perform essentially the same type of analysis to obtain the filter parameters. Although there may be speech quality differences depending, for example, on whether the covariance, autocorrelation or lattice method of linear prediction is used, these differences tend not be of a major nature. The upper bound on the vocoded speech quality is basically a function of the type of excitation used. This is discussed below for each of the three types of LPC vocoders.

Residual Excited Vocoder. In this type of vocoder [4], the residual signal is used to excite a filter that is the exact inverse of the filter used to generate the residual signal from the speech signal. Assuming no quantization errors in either the residual signal or the filter parameters, the synthesized signal s'(n) will be almost identical to the original signal s(n). Therefore, here, the analysis itself does not degrade the speech quality.

Voice Excited Vocoder. In this type of vocoder [5,6], a down sampled baseband comprises the source information. At the receiver the baseband is nonlinearly processed to obtain an excitation function with a flat spectrum. Even under no parameter quantization, the synthesized signal s'(n) will be different from s(n). Therefore, the speech model employed is already responsible for a certain change

in the speech quality when compared to the original. One method of estimating this change in quality would be to compare the filter excitation signal for this vocoder to the residual signal used in the residual excited vocoder. Such comparison is probably not straightforward, but it is made easier by the fact that the two signals are more or less time-synchronized (in terms of where pitch periods are, etc.).

Pitch Excited Vocoder. In this case, the excitation is either a sequence of pitch pulses or white noise. Here, s'(n) resembles s(n) in its gross features, but certainly not in the detailed signal structure. Also, unlike the voice excited case, s'(n) is generally not synchronized with s(n), because the voiced/unvoiced (V/UV) excitation is not synchronized with the residual signal, which makes it difficult to get an objective estimate of the change in quality due to the pitch excited model. This is unfortunate considering that the V/UV decision is perhaps the single most important one that affects the quality of s'(n). There are currently no established procedures for the automatic evaluation of V/UV decisions. The existing procedures are manual, in that intervention by a human is necessary to establish whether a voiced or an unvoiced decision would be appropriate for each frame in the analysis (and whether the extracted pitch value is accurate). In certain critical situations, such decisions are made by trial and error as to

which sounds better. There are other cases where a mixed voiced-frication source is more appropriate. Thus far, these cases have not been dealt with successfully in vocoders.

Because of the dearth of good testing procedures to evaluate the effects of the excitation on speech quality, we have decided to table this problem in our initial search for objective measures of quality.

Synthesis

Although a large part of the synthesis process is dictated by the type of model used and signal analysis performed, there remain a number of design choices in the synthesizer that can noticeably affect the synthesized speech quality. The major choices relate to the updating and interpolation of filter parameters, as well as the choice of the filter implementation structure. For example, we have found that if the analysis is performed time-synchronously, it is best to interpolate and update filter parameters time-synchronously as well [7].

Although there are important issues relating to filter implementation structure (for example, placing the gain at the output of a normalized filter [8] causes "clicks" to occur during large changes in gain), it is always possible to choose the implementation structure in such a way that

the structure itself contributes negligibly to the degradation of the quality.

Encoding

We include in this component

(1) the choice of the number of parameters to transmit,
(2) how to quantize them, and
(3) when to transmit them.

The parameters include the source parameters (the residual signal in a residual excited vocoder, or pitch and gain in a pitch excited vocoder), and the synthesizer parameters, which can take different forms, with the most popular being the log area ratios in an LPC vocoder [9], or the output energies of the channel filters in a channel vocoder. The choice of the number of parameters, along with their quantization, determine to a large extent the static signal quality at specific time instances, while the transmission and update rate determine the dynamic signal fidelity.

Conclusion

For narrow-band vocoder systems (less than 5000 bps), the encoder, as we have defined it, is the major determiner of speech quality. This is due to the heavy quantization that is necessary to produce low bit rates. Design issues in the analysis and synthesis are important, but for low rate systems, the encoder plays the major role.

4. General Framework

It follows from the previous section that, if the bulk of the synthesized speech quality is determined by the encoder, then one should be able to obtain an approximate objective measure of the quality difference between the original and vocoded speech by somehow comparing the parameter values at the input and output of the encoder. One could also include the interpolation in the synthesis component, and compare the parameter values at the synthesizer with the parameters at the input to the encoder (which are produced by the analysis). In any case, the problem is thus reduced from comparing the quality of two speech signals to comparing two sets of parameters that are related to each other in a well specified manner. This, in turn, implies that such comparisons or quality measuring procedures are to be built "inside" the vocoder instead of outside it. Comparisons are made between the unquantized parameters (reference system) and the parameter values used at the synthesizer (test system).

Inherent in the above analysis is that speech synthesized using the input parameters to the encoder is of very good quality. This is not difficult to achieve. For example, in an LPC vocoder, if the signal bandwidth is 5 kHz, then a 14-pole analysis every 10 ms would give unquantized parameters, which when used in the synthesis,

would result in speech whose quality is very good compared to the original speech. This does not necessarily mean that the encoder has to quantize the 14 filter parameters and transmit them every 10 ms. The restriction is merely on the analysis. The encoder may then choose a smaller (and perhaps variable) order for transmission, and at a lower (and perhaps variable) rate [7].

We now state the three observations (assumptions) that form the basis for our work in developing objective quality evaluation measures:

(1) Speech synthesized from unquantized parameters, extracted every 10 ms, is of very good quality compared to the original speech.
(2) Except for pitch and gain, the fidelity of the short-time spectrum is the principal determiner of quality.
(3) The spectrum is uniquely defined by the filter parameters.

The first observation gives us an anchor point defined in terms of the system parameters and against which to compare quantized realizations of the same utterance. The second and third observations relate the filter parameters to speech quality through the concept of spectral fidelity. This, then, gives us a framework within which to develop the desired objective measures of speech quality.

5. An Initial Experiment

Given a speech utterance processed by an LPC vocoder, an objective measure summarizes the error or deviation between the reference and the test sets of parameters in terms of a single number which we shall call an objective evaluation score. The objective score would be expected to reflect the perceived quality (relative to the reference) of the speech utterance if, indeed, the objective measures were sensitive to all quality-determining factors. It is unreasonable, and perhaps too simplistic, to expect that one objective measure could always correctly predict perceived speech quality. The chance of such a prediction may be enhanced by combining a number of objective measures in some fashion to obtain an overall objective score. Each measure may be sensitive to some aspects of quality. Ultimately, we plan to perform a multidimensional analysis [11] on the objective scores obtained from a number of measures with the hope of relating them to different quality dimensions yet to be discovered. For the present study, however, we chose to develop a number of objective measures and investigate each of them separately so as to become familiar with their properties.

For each data frame, an error between the reference and the test parameters is computed using an appropriate "distance" measure. Ideally, such frame errors should be

computed only at selected points in time within the speech utterance that are "perceptually significant." For the purposes of the present study, we simply computed the frame error at a fixed rate, say, every 10 ms. We thus had two problems. (1) To develop suitable distance measures to compute frame errors. (2) To combine all the frame errors within a speech utterance into one number, which provides the objective score.

We considered several distance measures for computing the error between the reference and test parameters of a given frame. These measures were based on the power spectrum of the all-pole linear prediction filter. Traditional mean squared differences between log spectra, as well as other measures were used. The errors were also frequency weighted in different ways, including a weighting based on the articulation index [10]. The resulting error sequence at each frame was then combined to give the overall objective score. The sequence was time weighted using the filter gain and the "spectral difference" (rate of change of spectrum) between frames.

These objective measures were used in an initial experiment to correlate the objective scores with the results of a rank ordering experiment of subjective quality that compared different vocoder systems [11]. Different combinations of objective measures were used in the

experiment. Comparisons of the objective and subjective scores indicated that no single objective measure was able to always predict correctly the subjective rank ordering of vocoded speech utterances. Furthermore, the objective scores were heavily clustered (relative to the subjective scores) for the different systems, indicating a lack of separability.

6. Program for Research

Based on our initial experiments it became clear that what we need is a step-by-step program to understand the different aspects of the relations between spectral variations and speech quality, in order to be able to begin developing the desired objective measures of quality. First, we shall attempt to discover the quality determining factors in the spectrum independent of time. Following that, we shall attack the more difficult problem of discovering the time-dependent quality determining factors.

As a first step, we have begun to develop spectral distance measures that are consistent with published perceptual data on vowel difference limens. This work is described in a separate paper [10]. One of the important conclusions there is that traditional distance measures between log spectra are not consistent with perceptual data.

7. Conclusions

In this paper we presented the rationale behind the general framework for the objective evaluation of vocoder speech quality. The framework calls for inserting these objective measures inside the vocoder to compare the sets of filter parameters after analysis and before synthesis, in order to observe the effects of encoding and interpolation on the resulting spectra. Spectral variations, in turn, are related to speech quality. A step-by-step program has been initiated to discover the time-independent as well as time-dependent quality determining factors in the short-time spectrum.

References

1. L.L. Beranek, Acoustics, New York: McGraw-Hill, 1954.

2. K.D. Kryter, The Effects of Noise on Man, New York: Academic Press, 1970.

3. J.L. Flanagan, Speech Analysis Synthesis and Perception, Second Edition, New York: Springer-Verlag, 1972.

4. C.K. Un and D.T. Magill, "The Residual-Excited Linear Prediction Vocoder with Transmission Rate below 9.6 kbits/s," IEEE Trans. Comm., Vol. COM-23, 1466-1474, Dec. 1975.

5. C.J. Weinstein, "A Linear Prediction Vocoder with Voice Excitation," EASCON '75, Washington, D.C., 30A-30G, Sept. 29-Oct. 1, 1975.

6. B. Atal, M. Schroeder and V. Stover, "Voice-Excited Predictive Coding System for Low Bit-Rate Transmission of Speech," Int. Conf. Comm., San Francisco, Ca., June 1975.

7. J. Makhoul, R. Viswanathan, L. Cosell and W. Russell, Natural Communication with Computers, Final Report, Vol. II, Speech Compression Research at BBN, Report No. 2976, Bolt Beranek and Newman Inc., Cambridge, Mass., Dec. 1974.

8. A. Gray, Jr., and J. Markel, "A Normalized Digital Filter Structure," IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-23, 268-277, June 1975.

9. R. Viswanathan and J. Makhoul, "Quantization Properties of Transmission Parameters in Linear Predictive Systems," IEEE Trans. Acoustics, Speech and Signal Processing, Vol. ASSP-23, 309-321, June 1975.

10. R. Viswanathan, J. Makhoul and W. Russell, "Towards Perceptually Consistent Measures of Spectral Distance," IEEE Int. Conf. Acoustics, Speech and Signal Processing, Philadelphia, April 1976.

11. A.W.F. Huggins and R.S. Nickerson, "Some Effects of Speech Materials on Vocoder Quality Evaluations," J. Acoust. Soc. Am., Vol. 58, Supplement No. 1, S129, Fall 1975.

APPENDIX C

TOWARDS PERCEPTUALLY CONSISTENT MEASURES OF SPECTRAL DISTANCE

NSC Note 88, March 8, 1976

(Authors: R. Viswanathan, John Makhoul and W. Russell)

(This paper was presented at the 1976 International Conference on Acoustics, Speech and Signal Processing, Philadelphia, April 12-14, 1976)

TOWARDS PERCEPTUALLY CONSISTENT MEASURES OF SPECTRAL DISTANCE

This paper considers distance measures for determining the deviation between two smoothed short-time speech spectra. Since such distance measures are employed in speech processing applications that either involve or relate to human perceptual judgment, the effectiveness of these measures will be enhanced if they provide results consistent with human speech perception. As a first step, we suggest Flanagan's results on difference limens for formant frequencies as one basis for checking the perceptual consistency of a measure. A general necessary condition for perceptual consistency is derived for a class of spectral distance measures. A class of perceptually consistent measures obtained through experimental investigations is then described, and results obtained using one such measure under Flanagan's test conditions are presented.

1. Introduction

Given two smoothed short-time speech spectra, a fundamental problem in speech processing is to determine the distance or the amount of deviation between the two spectra. In speech recognition, the two spectra may correspond to two different speech sounds, or perhaps two different versions of the same sound [1-3]. In speaker verification or

identification, the two spectra may correspond to speech produced by either two different speakers or by the same speaker on two different occasions [4,5]. In variable frame rate speech compression, two adjacent analysis frames may have produced the two spectra [6,7]. In the problem of objective evaluation of vocoded speech quality, which the authors have recently formulated [8], the two spectra may correspond to the quantized and the unquantized sets of filter parameters. Still another application of spectral distance measures is in the spectral sensitivity analysis needed for optimal parameter quantization [9].

These examples clearly bring out the importance of spectral distance measures in speech processing. The extent to which a distance measure is valid greatly determines the efficiency of the underlying task in which it is employed. Inasmuch as one strives to achieve a machine performance that is close to what a human can do under the same situation (e.g., first two applications above), or inasmuch as the vocoded speech is to be perceived by human listeners, it is appropriate to require of these distance measures to be at least consistent with the known results of human perception. The work reported in this paper represents a first step towards obtaining perceptually consistent measures of spectral distance.

About two decades ago Flanagan reported perceptual results for determining difference limens for formant frequencies of vowels [10]. One of his results is particularly relevant to this paper. Briefly, when two formants are in close proximity, human perception exhibits an asymmetrical pattern in that moving one of the two formants closer to the other by a given amount produces a larger perceived quality difference than moving that formant away from the other by the same amount. On the other hand, the same formant shifts produce a symmetrical pattern when the two formants are well separated. We use this result as one basis for checking the perceptual consistency of spectral distance measures.

Smoothed spectra can be obtained by using a number of methods such as filter bank, cepstrum, and linear prediction (LP). For simplicity, we focus in this paper on LP spectra, although most of the discussions presented below apply to other types of spectra as well. The LP spectrum is given by [11]

$$P(\omega) = \frac{G^2}{S(\omega)} = \frac{R_o V_p}{\left| 1 + \sum_{k=1}^{p} a_k e^{-j\omega k} \right|^2} , \qquad (1)$$

where G is the linear predictor gain, R_o is the speech signal energy, V_p is the normalized prediction error, $S(\omega)$ is the spectrum of the inverse filter and a_k, $1 \leq k \leq p$, are the predictor coefficients.

2. Spectral Distance Measures

Let $d(X,Y)$ denote the distance or deviation between the spectra $X(\omega)$ and $Y(\omega)$. From a mathematical viewpoint, one may be tempted to insist that the distance measure satisfy the three axioms of a metric:

(a) Positive definiteness: $d(X,Y) \geq 0$, $d(X,Y)=0$ iff $X=Y$;
(b) Symmetry: $d(X,Y)=d(Y,X)$;
(c) Triangle inequality: $d(X,Y) \leq d(X,Z)+d(Z,Y)$.

We require, however, only the property (a) to be true. There are many examples in real life where distance symmetry does not hold. There is no evidence to support the validity of a symmetrical distance in the context of human speech perception. For a similar reason, we do not insist that the property (c) be necessarily true. We postulate that if a distance measure is perceptually consistent, it will prove to perform better in applications involving, or relating to, human perception.

Normalization

Before we define a measure of distance between two LP spectra $P_1(\omega)$ and $P_2(\omega)$, it may be desirable to normalize these spectra in some fashion. For instance, they may be normalized to have the same arithmetic mean (AM) or total energy. Alternately, they may be normalized to have the same geometric mean (GM), i.e., the log spectra will have the same average.

Error Definition

An error function between the normalized spectra can be defined either in the (linear) spectral domain as

$$e(\omega) = P_1(\omega) - P_2(\omega) \quad , \tag{2}$$

or, in the log spectral domain as

$$e(\omega) = \log P_1(\omega) - \log P_2(\omega) \quad . \tag{3}$$

Other reasonable error definitions include

$$e(\omega) = [P_1(\omega) - P_2(\omega)]/P_1(\omega) \quad , \tag{4}$$

$$e(\omega) = P_1(\omega)/P_2(\omega) \quad . \tag{5}$$

Spectral Distance Measure

A large class of spectral distance measures can be defined as the weighted L_k norm:

$$d_k(P_1,P_2,W) = \left[\frac{\int_{-\pi}^{\pi} W(P_1(\omega),P_2(\omega),\omega)\,|e(\omega)|^k d\omega}{\int_{-\pi}^{\pi} W(P_1(\omega),P_2(\omega),\omega)\,d\omega} \right]^{\frac{1}{k}} \tag{6}$$

where the weighting function W in general depends on $P_1(\omega)$, $P_2(\omega)$ and frequency ω, and takes only positive values. If the error is defined as in (4) or (5), the distance measure in (6) is not symmetric. Also, if the weighting function depends explicitly on P_1 and P_2, the resulting distance measure is in general not symmetric. In all other cases, a

symmetric distance measure results. In the absence of any weighting, d_{-1} is the harmonic mean, d_0 is the GM, d_1 is the AM, and d_2 is the root mean square value of the absolute error function. Between the minimum $d_{-\infty}=\text{Min}|e(\omega)|$ and the maximum $d_{\infty}=\text{Max}|e(\omega)|$, d_k is a monotonically increasing function of k.

The weighting function W in (6) is used to differentially weight individual errors and is determined based on some concept of speech perception. Notice that any constant multiplicative factor in the weighting function does not affect the distance measure. Some specific weighting functions are discussed in Section 4.

Examples: References [2,6,7] use d_1 with the error defined as in (5). (With a Gaussian assumption, this measure becomes a likelihood ratio [2].) Reference [9] employs d_1 with the error given by (3). Cepstral distance measures used in [1,4] have been shown to be highly correlated to d_2 with the error as in (3) [12].

3. A Necessary Condition for Perceptual Consistency

Fig. 1 shows two plots of spectral deviation or distance versus frequency shift of the second formant causing that spectral deviation. (Frequencies of the other three fixed formants and the nominal value of the second formant frequency are given in the figure. Fixed bandwidths

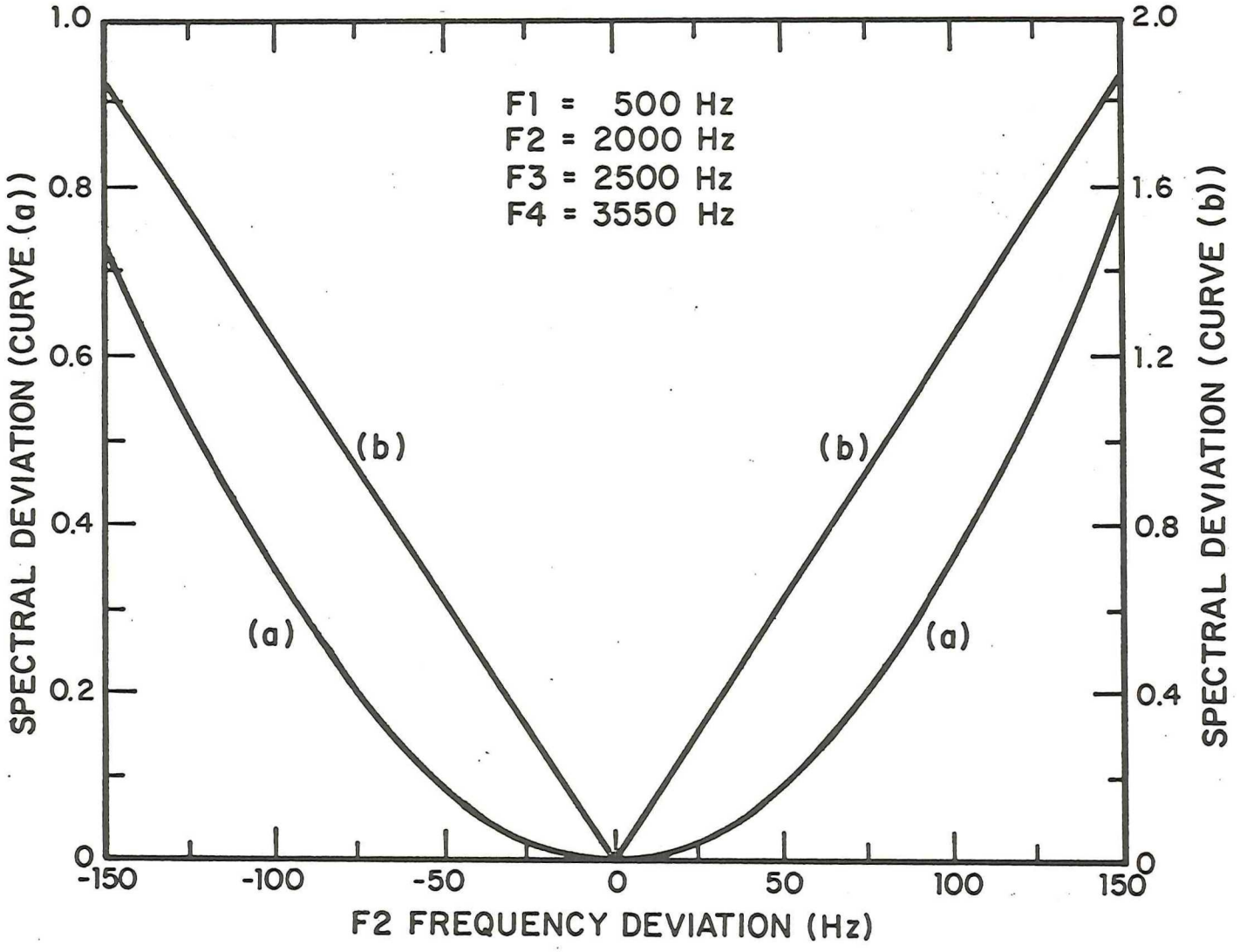


Fig. 1. Plots of spectral deviation versus shift in second formant frequency about 2000 Hz for two spectral distance definitions.

of all the formants are as in [10].) Fig. 1(a) corresponds to the error definition (5) while Fig. 1(b) corresponds to the error definition (3). Both plots were obtained using GM normalization, k=1 and no weighting in (6). (We have plotted log d for plot (b) so that ordinates of both plots are in decibels.) The almost symmetrical plots in Fig. 1 do not conform with properties given by Flanagan (see Fig. 4(c) in [10]).

Notice that the two distance measures that produced the plots in Fig. 1 depend only on the ratio of the spectra P_1 and P_2 (in view of (3) and (5)). Below we prove that with GM normalization, any distance measure which is a function of only the ratio of the spectra is necessarily perceptually inconsistent. First, we give our working definition of perceptual consistency, based on Flanagan's results [10].

Working Definition of Perceptual Consistency: Let X and Y be two vowel spectra, such that Y is identical to X except that one of the formant frequencies F is shifted by a variable amount F. A given spectral distance measure d(X,Y) between X and Y is said to be perceptually consistent if

(a) when F is close to another formant F', d(X,Y) exhibits asymmetry such that it is greater when F is moved ΔF towards F', than when F is moved ΔF away from F';
(b) such asymmetry decreases as F and F' are further apart.

Now, consider a class D_g of spectral distance measures defined by (6) where the error $e(\omega)$ is computed after GM normalization of the spectra. For this class of distance measures, a necessary condition for perceptual consistency is provided below in the form of a theorem.

Theorem: A necessary condition for any spectral distance measure $d(P_1,P_2)$ in the class D_g to be perceptually consistent (as defined above) is that it not be a function of only the ratio of the two spectra P_1 and P_2.

Proof: Assume that a distance measure in D_g violates the necessary condition. We show that this distance measure is not perceptually consistent. Let P_2 be obtained from P_1 by shifting only one of its formant frequencies while keeping all other parameters intact. Let the denominator $S(\omega)$ in (1) be factored into $R(\omega)$ and $S'(\omega)$, where $R(\omega)$ is the contribution to the spectrum from the formant under consideration and $S'(\omega)$ represents the contributions from all other poles of the linear predictor. Thus, $P_1(\omega) = 1/(R_1(\omega).S'_1(\omega))$ and $P_2(\omega) = 1/(R_2(\omega).S'_1(\omega))$, where $R_2(\omega)$ is the perturbed version of $R_1(\omega)$. This gives the result that the ratio of P_1 and P_2 depends only on the formant under consideration. Specifically, the ratio does not depend on whether or not this formant is in close proximity to another formant. This clearly establishes that the measure is not perceptually consistent according to our working definition.

With other types of spectral normalization, the ratio of gain terms (G^2 in (1)) of the two spectra depends in general on the overall shape of the spectrum. For instance, with AM normalization, this ratio is between the normalized prediction errors (V_p in (1)) corresponding to the two spectra, which depend on the total spectral shapes [3]. Establishing a general necessary condition for perceptual consistency in these cases is difficult. However, with AM normalization, our experimental results show that when the necessary condition stated above is violated, perceptual consistency is not obtained.

We do not wish to state that perceptually inconsistent measures are not useful. In fact, in the applications mentioned in the introduction, many such measures have been succesfully used. We suggest, however, that use of perceptually consistent measures in these applications may lead to an improved performance of the underlying tasks.

4. Weighting Functions

We have investigated a number of reasonable frequency weighting functions [13]. A brief discussion of some of these weighting functions is given below.

Spectral Intensity Weighting

Since formant peaks of a spectrum are perceptually important, it is reasonable to emphasize spectral errors that occur close to formant peaks. One way of achieving this error weighting is to use $P_1(\omega)$, $P_2(\omega)$ or some generalized mean of the two as weighting functions.

Frequency Derivative Weighting

An alternate method of emphasizing spectral errors that occur close to formant peaks is to employ a suitable function of first and second derivatives of $P_1(\omega)$ or $P_2(\omega)$ for weighting the errors.

Articulation-Index (AI) Based Weighting

AI is a physical measure that is highly correlated with subjective speech intelligibility results. Since it is not unrealistic to consider speech intelligibility and quality as related phenomena, we have derived, by adapting some of the results used in AI computation, a weighting function which decreases exponentially with frequency: $W=\exp(-\alpha\omega)$, where α is a particular constant [13].

All the spectral distance measures that we investigated, even with the use of the above weighting functions, had one common problem in that for the case when the first formant frequency was shifted about the nominal

value of 300 Hz, a given amount of left shift always produced a larger spectral deviation than a right shift of the same amount, which is just the opposite of what Flanagan reported (see Fig. 3(a) in [10]). (We found, however, that some of these measures and weighting functions produced the right types of asymmetry in other test conditions considered by Flanagan.) To attempt to overcome this problem, we investigated the following weighting functions based on perceived loudness.

Perceived Loudness Weighting

Based on the work of Stevens [14], we define the perceived loudness function $L(\omega)$ of a spectrum $P(\omega)$ as $[P(\omega)A(\omega)]^{1/3}$, where $A(\omega)$ is shown plotted in Fig. 2. Notice the sharp change of $A(\omega)$ at low frequencies, which may be used to our advantage to overcome the problem mentioned above. The weighting function may then be defined in terms of $A(\omega)$ or $L(\omega)$. We have investigated the following weighting functions: $W=A(\omega)$; $W=L_1(\omega)$ (perceived loudness of $P_1(\omega)$); $W=L_2(\omega)$; $W=|L_1(\omega)-L_2(\omega)|$. Only the weighting function $W=A(\omega)$ produced the right asymmetry for the case when the first formant frequency was shifted about its nominal value of 300 Hz.

In the next section, we give examples of perceptually consistent distance measures which use the weighting function $A(\omega)$.

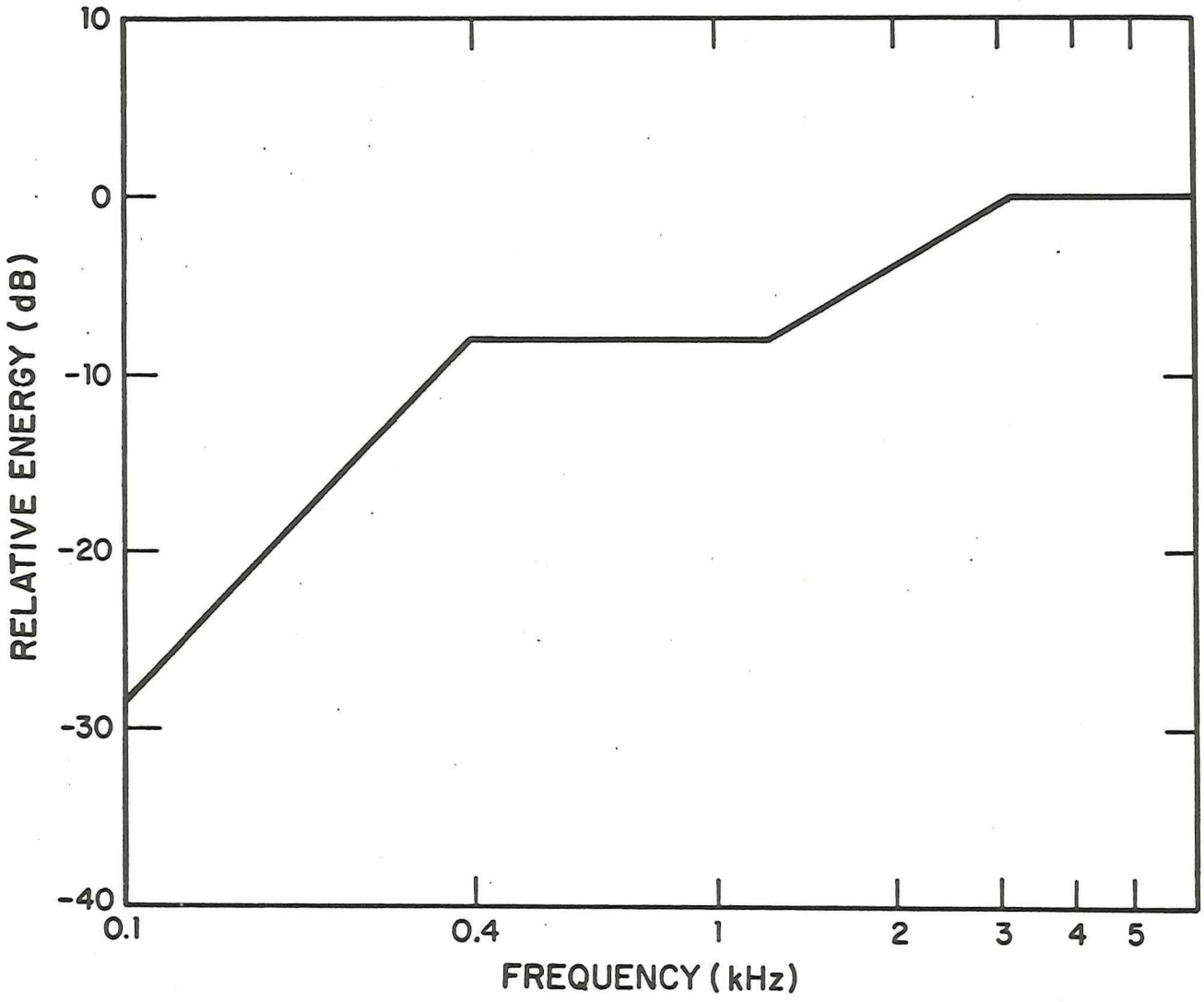


Fig. 2. Frequency weighting function $A(\omega)$.

5. A Class of Perceptually Consistent Distance Measures

Our experimental investigations have led to a class of spectral distance measures which produce the right types of asymmetry attributable to formant interaction under all test conditions considered by Flanagan. This class is defined by (6) with GM normalization, the spectral error defined in the (linear) spectral domain as in (2), and the weighting function $A(\omega)$ shown in Fig. 2.

Figs. 3-5 show plots of spectral distance versus formant frequency shift under three different test conditions for the above measure with $k=1$ in (6). These plots compare rather nicely to the corresponding ones that Flanagan has given. Notice that while our spectral distance plots in general have a monotonically increasing tendency, Flanagan's plots reach a constant 100 for large formant frequency shifts due to the fact that subjects in his tests were asked to merely say if they perceived the two speech sounds corresponding to unperturbed and perturbed sets of formants as being different rather than to quantify the amount of quality difference they perceived between the two sounds.

The effectiveness of the weighting $A(\omega)$ is particularly apparent in the low frequency region. Fig. 6 shows plots of spectral distance with and without this weighting, other conditions being the same, for the case when the first formant is shifted about 300 Hz. The unweighted measure

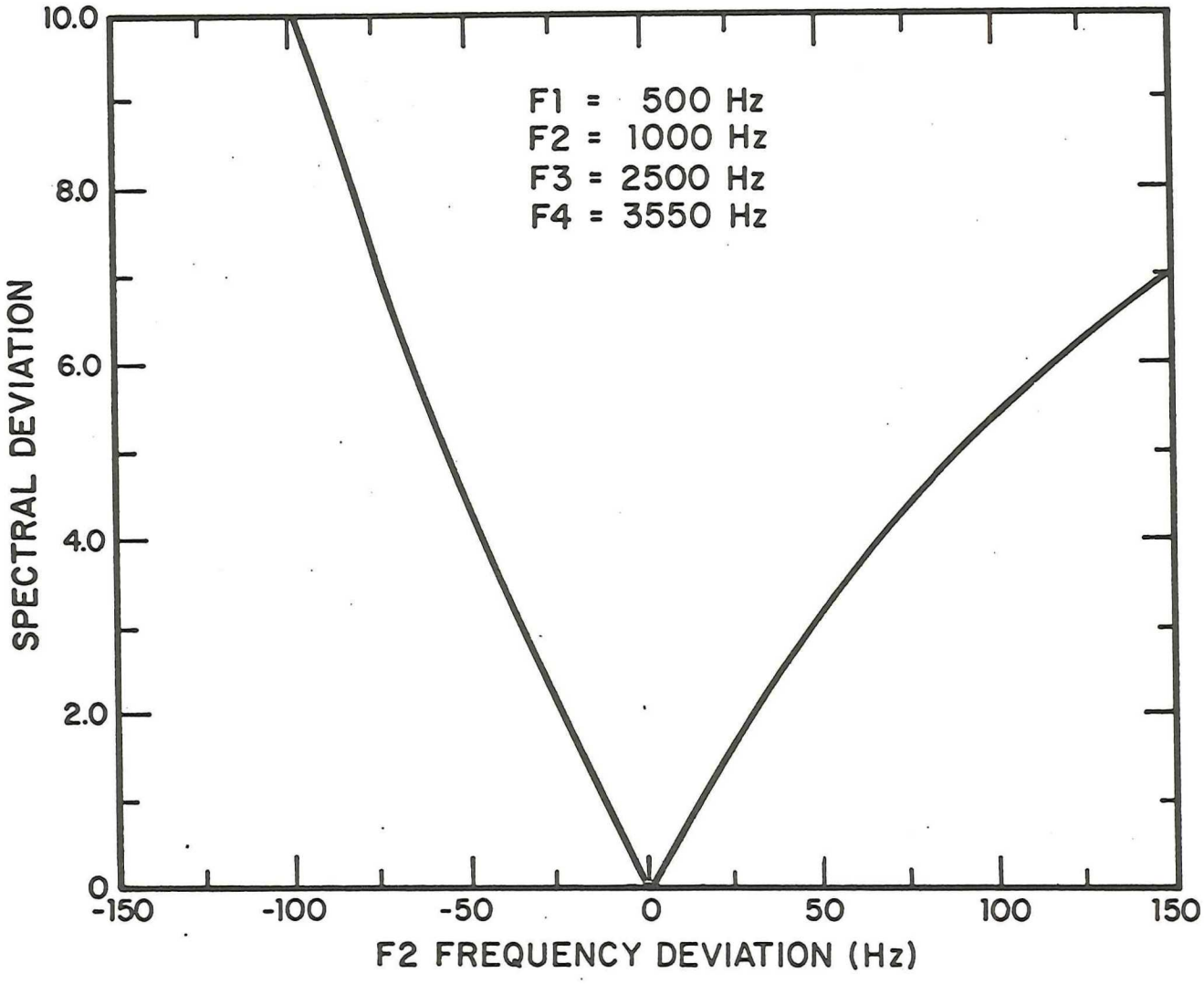


Fig. 3. Spectral deviation versus shift in second formant frequency about 1000 Hz for the spectral distance $d_1(P_1,P_2,A)$ in (6), where P_1 and P_2 are GM-normalized spectra, $e(\omega)$ is given by (2) and $A(\omega)$ is shown in Fig. 2.

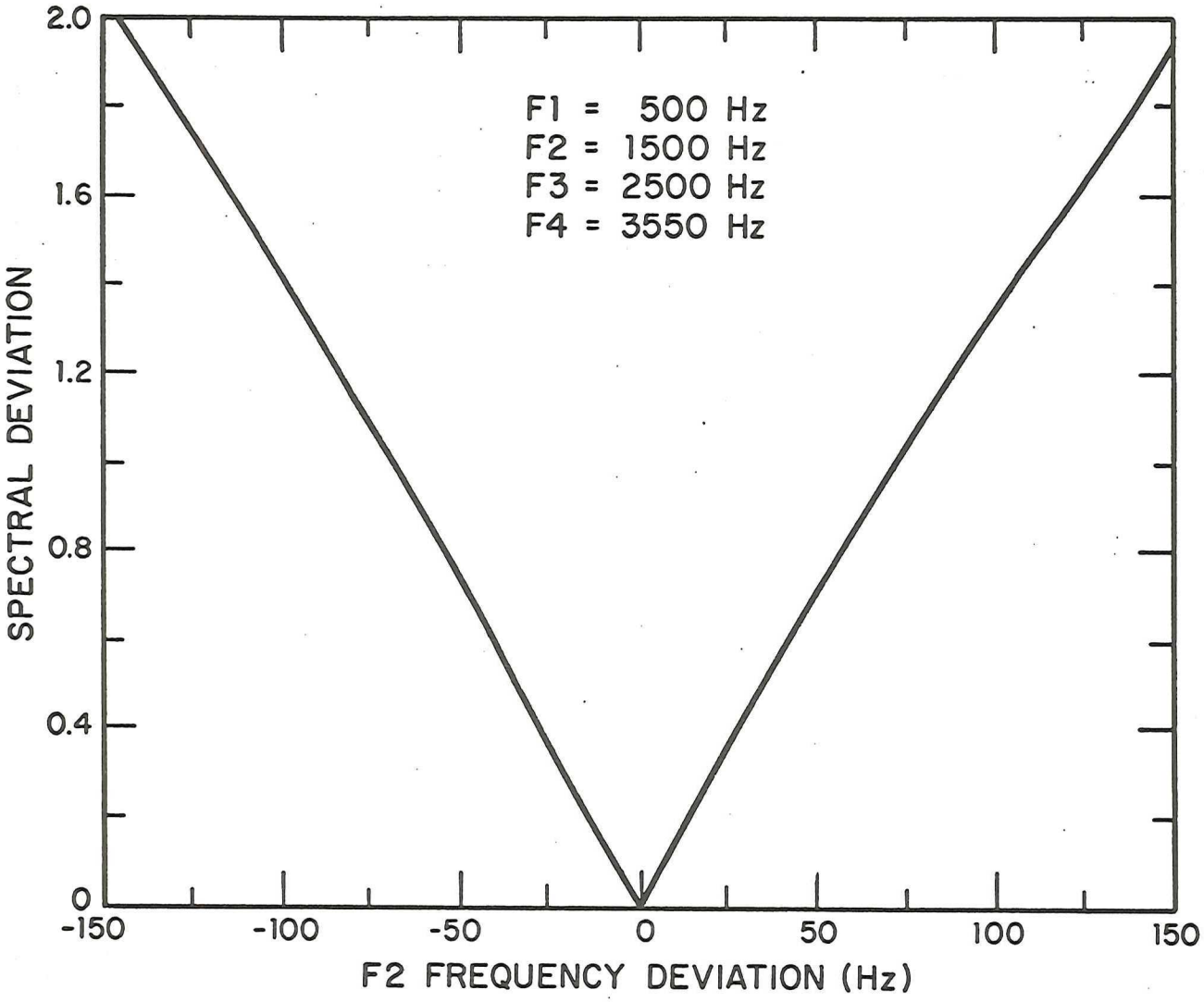


Fig. 4. Spectral deviation versus shift in second formant frequency about 1500 Hz for the spectral distance $d_1(P_1,P_2,A)$.

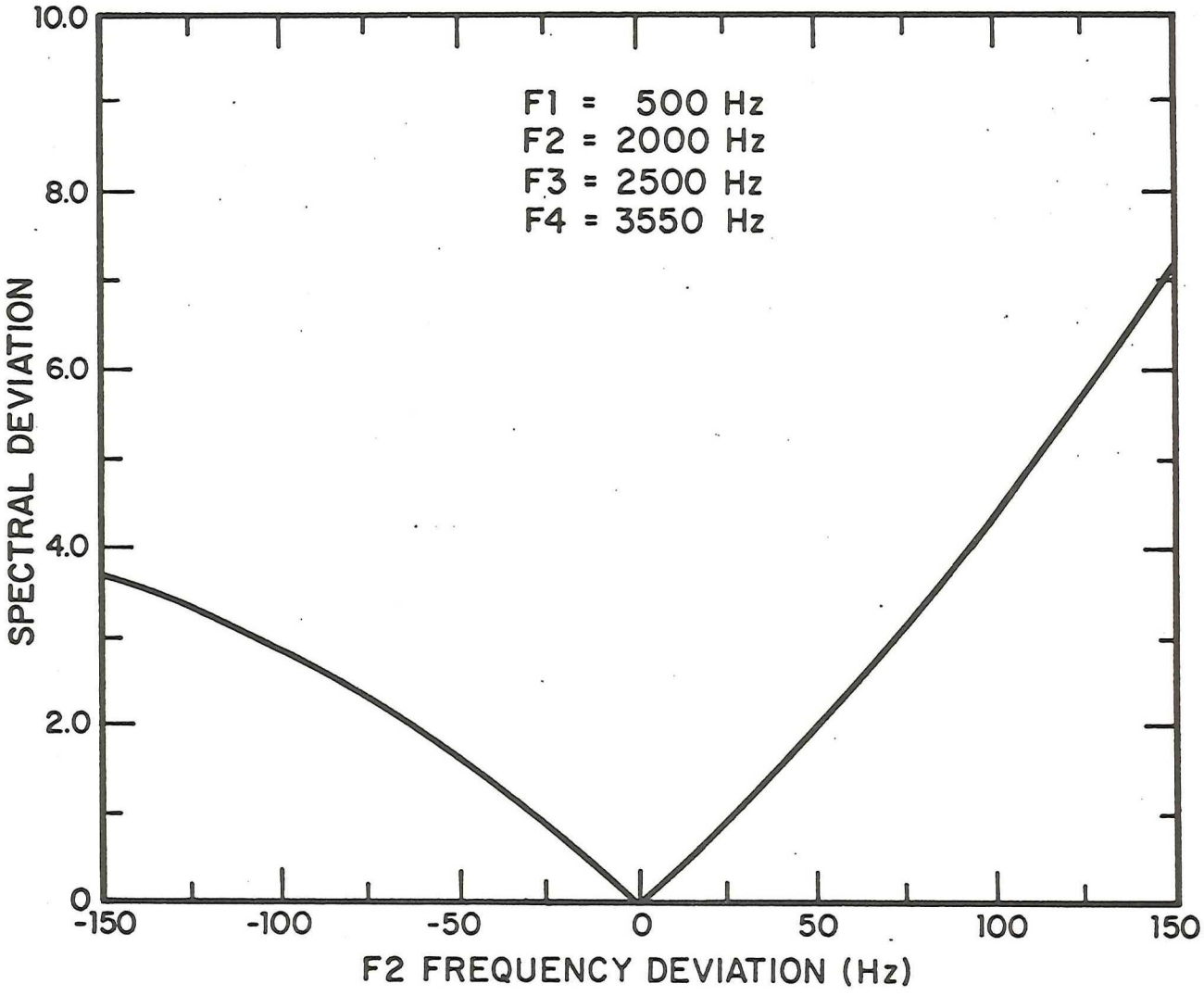


Fig. 5. Spectral deviation versus shift in second formant frequency about 2000 Hz for the spectral distance $d_1(P_1,P_2,A)$.

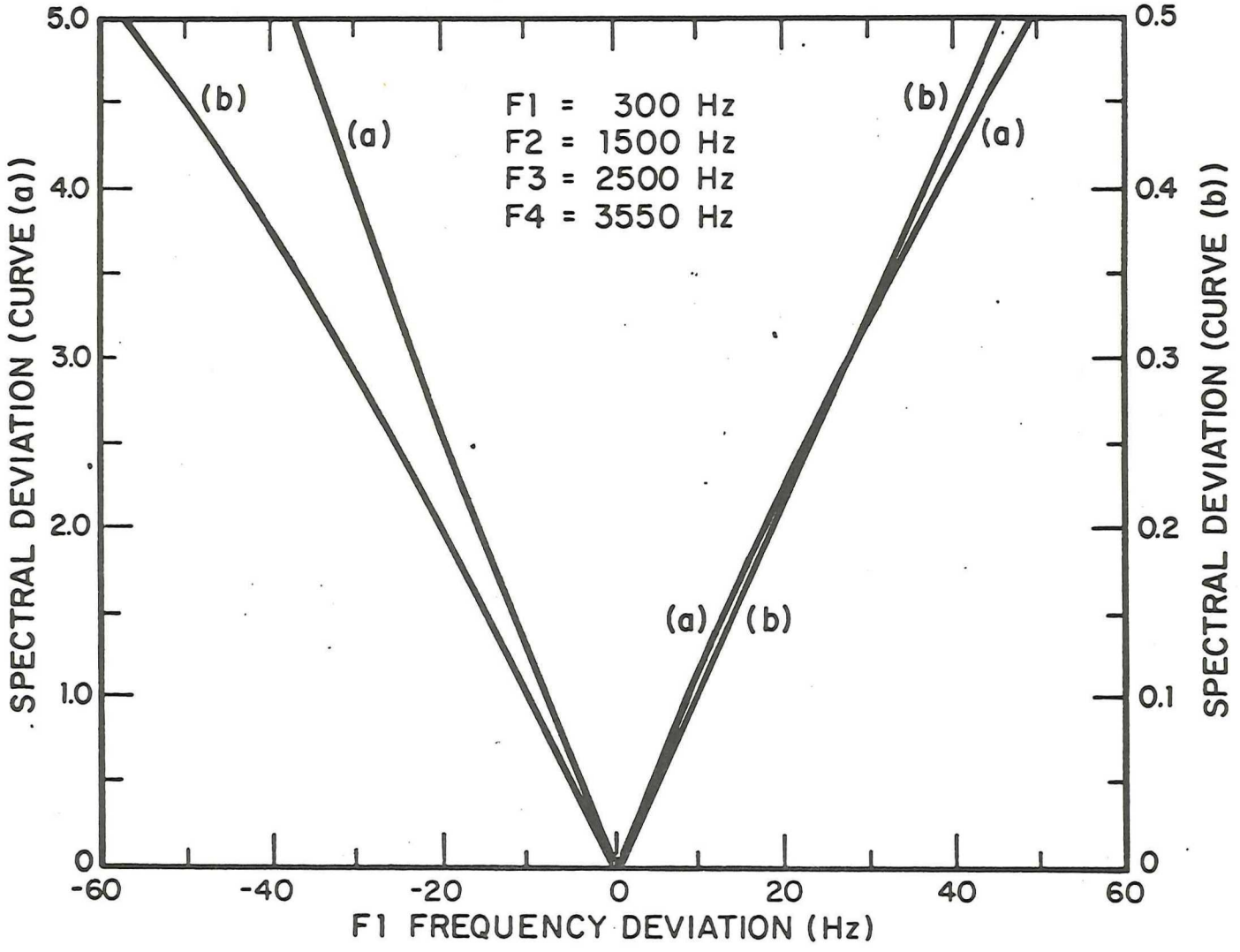


Fig. 6. Plots of spectral deviation versus shift in first formant frequency about 300 Hz for the distance measure d_1 between GM-normalized spectra (a) without any frequency weighting, and (b) with frequency weighting by $A(\omega)$.

gives a slight asymmetry but in the wrong sense, Fig. 6(a), while the weighted measure produces the right asymmetry as shown by Fig. 6(b).

A disadvantage of the distance measures presented in this section is that they are dependent on the energy of the spectra. (Notice that distance measures which are functions of only the ratio of spectra do not suffer from this disadvantage.) With energy dependent measures, comparison of spectral distances obtained, for instance, for different analysis situations can be meaningfully done only after suitably scaling the distance values. A reasonable condition to impose on such scaling is that the spectral distances corresponding to the formant frequency difference limens at the different frequencies be approximately equal. This will be our next step in refining the class of perceptually consistent spectral distance measures that we suggested above.

6. Conclusions

We have reported preliminary results of an ongoing work on perceptually consistent spectral distance measures. Our experience has been that GM normalization works better than AM normalization inasmuch as one is looking for sensitivity to interaction of formants. The results we have presented in this paper show that the distance is best defined in terms of the difference in the (linear) spectral values.

Besides continuing our investigation reported here, we plan to use the developed measures in several applications.

References

1. A. Ichikawa, Y. Nakano and K. Nakata, "Evaluation of Various Parameter Sets in Spoken Digits Recognition," IEEE Trans. AU, 202-209, June 1973.
2. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. ASSP, 67-72, Feb. 1975.
3. J. Makhoul,"Linear Prediction in Automatic Speech Recognition," in Speech Recognition, R. Reddy (Ed.), New York: Academic Press, 1975.
4. B.S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification," JASA, 1304-1312, June 1974.
5. S.F. Boll, "Waveform Comparison Using the Linear Prediction Residual," Comp. Science Dept., Univ. Utah, 1975.
6. D. T. Magill,"Adaptive Speech Compression for Packet Communication Systems," Proc. Nat'l Telecommun. Conf., Nov. 1973.
7. J. Makhoul, R. Viswanathan, L. Cosell and W. Russell, Natural Communication with Computers, Vol. II, Speech Compression Research at BBN, Rept.No. 2976, Dec. 1974.
8. J. Makhoul, R. Viswanathan and W. Russell, "A Framework for the Objective Evaluation of Vocoder Speech Quality," Internat'l Conf. ASSP, Philadelphia, April 1976.
9. R. Viswanathan and J. Makhoul, "Quantization Properties of Transmission Parameters in Linear Predictive Systems," IEEE Trans. ASSP, 309-321, June 1975.
10. J.L. Flanagan, "A Difference Limen for Vowel Formant Frequency," JASA, 613-617, May 1955.
11. J. Makhoul, "Linear Prediction: A Tutorial Review," Proc. IEEE, 561-580, April 1975.
12. A.H. Gray, Jr. and J.D. Markel, "Distance Measures for Speech Processing," Submitted for publication in IEEE Trans. ASSP.
13. BBN Quarterly Progress Report, Command and Control Related Computer Technology, BBN Rept. No. 3122, Sept. 1975.
14. S.S. Stevens, "Perceived Level of Noise by Mark VII and Decibels (E)," JASA, 575-600, Feb. 1972.